CHARACTERISING UNDERSTANDING & THE UNDERSTANDING SUBJECT

-Sven Delarivière -

VUB VRIJE UNIVERSITEIT BRUSSEL

Dissertation submitted for the degree of Doctor of Philosophy and Moral Sciences.

CHARACTERISING UNDERSTANDING & THE UNDERSTANDING SUBJECT

Sven Delarivière

Dissertation submitted for the degree of Doctor of Philosophy and Moral Sciences.

Members of the doctoral jury

Bart Van Kerkhove (Vrije Universiteit Brussel, promotor)
Jean Paul Van Bendegem (Vrije Universiteit Brussel, co-promotor)
Line Edslev Andersen (Aarhus University)
Henk Willem de Regt (Radboud University Nijmegen)
Karen Francois (Vrije Universiteit Brussel)
Joachim Frans, (Vrije Universiteit Brussel)

The investigations reported in this dissertation have been supported by
an Aspirant Fellowship from the Fonds voor Wetenschappelijk Onderzoek (FWO).

(2020)

Acknowledgements

In spirit of the claims I make in the later chapters, I should acknowledge that my credit for this dissertation does not preclude additional credit being due to the many people and processes that helped me along the way.

Without my parents, I would never have been burdened with the blessing of existence, been given the the middle-class privilege to be financially sponsored all through a Philosophy degree (even after attaining an Arts degree) or even been allowed to pursue a degree in something that is labelled as "unproductive to society" by the ruling class (wait till they find out about stockholders, landlords and lobbyists – oh wait). Without being born a straight, white, cis-male, I would never have belonged to the group that, unfortunately, finds it disproportionally easier to pursue their interests. Without natural selection and cultural evolution, I would never even have been suited to, nor interested in, something as peripheral to physical survival and procreation as philosophy. Without the many contributions of philosophers past and present, I would never have had shoulders of giants to stand on. Without my colleagues in the CLPS, I would have found the research-process less supportive, stimulating and silly. Without my supervisor, I would never have considered applying for a PhD, nor found my research-process as unconstrained and smooth.

Without my dearest partner, Flora, I would have never made it to the other side as wisely, healthily or happily.

Whatever is left is all me.

DISSERTATION

Table of Contents

PART II CHARACTERISING EPISTEMIC SUBJECTS

DISSERTATION

Introduction

Humans strive for understanding. Understanding is a valued aim or trait in any discipline, and it is fair to say that we sometimes uncontroversially attribute that trait to certain subjects. But is there a systematic way to reveal what understanding-attributions are or should entail? What is it in those situations of uncontroversial attribution (or dismissal) that guides us - or should guide us - in ascribing (or denying) understanding? Answering this question involves providing a conceptual characterisation of "understanding" that is explanatory as well as philosophically coherent and consistent, but which furthermore allows us to explain who does and does not understand, as well as why or why not. The latter is especially is especially relevant when it comes to assessing unconventional (and therefore unintuitive) cases of understanding such as extended, collective or artificial understanding.

There are many questions involved: What is understanding? How does it manifest itself? What makes it *about* something? Are the manifestations different depending on what it is about? What do we value about it? Who can it be attributed to? What makes for good understanding? How do we evaluate it? Are there degrees of understanding? Is there such a thing as complete understanding? Are there different kinds? Are the normative standards atemporal or universal? How stable must understanding be? Does it matter how we come to understand? Must understanding be internally consistent? How do we guarantee this? Does understanding involve consciousness? Can someone understand on the basis of following rules, idealisations, false beliefs? Does the use of tools discredit understanding? Can groups understand? Can we impart understanding onto computers?

I will, in this dissertation, set up my conceptual characterisation of "understanding" and "the understanding subject." This is an act of much needed conceptual clarification. If there is any field for which this need is most urgently felt, it is that of epistemology. But the concept of understanding encompasses many aspects (e.g. the acts, abilities, degrees, scope, quality and evaluation of understanding, the intelligence, minds, beliefs, efficiency, potential and resources of an understander, the accuracy, scientificness, contextuality and objects of understanding), stretching into many different fields (e.g. epistemology, philosophy of science, philosophy of mathematics, philosophy of action, philosophy of mind, cognitive science and even computer science). It is difficult to address the notion of understanding, since it involves many different aspects that stretch different fields of philosophy. There is some further patchiness stemming from the different perspectives on the

different objects of understanding.[1] Therefore, I believe the characterisation of understanding and its subject would benefit from a big picture approach. Not one which tries to solve the next problem within a particular field with a particular focus, but one which can tie together several insights from a variety of fields regarding the many aspects of understanding.

Recent discussions in epistemology have started to take seriously the question whether some epistemic properties may be suitably attributed (to a lesser or equal extent) to entities other than human individuals, such as groups (collective epistemology), artefacts (android epistemology) or individuals in conjunction with "external" components (extended epistemology). The rising debate in epistemology shares similarities with developments from cognitive science (e.g. distributed, extended, and artificial cognition) and philosophy of mind (e.g. theories of agency, collective intentionality, active externalism and emergence), where there is already a widespread literature and several developed frameworks and analytic toolboxes. But so far these developments have only to a very limited extent influenced the epistemology literature. Furthermore, the literature is fragmented based on the properties and entities they discuss. Focusing on epistemic subjecthood, generally, would force us to consider what unites the demarcation principles at work in the particulars as well as what the explanatory role of epistemic subjecthood amounts to and how it fares in explaining entities beyond human individuals.

Considering the scarce and disjointed nature of the present literature on understanding, it will be a fruitful and much-needed step to string together claims of epistemology and other relevant fields in a clear and systematic big picture approach where we can keep track of where particular claims fit in and what they have bearing on. This dissertation is an attempt to present such an approach. Although I am primarily writing with epistemologists in mind (the field where this kind of conceptual clarification is most at home), my research does not fit into one discipline. It is located between the cracks of epistemology, philosophy of mind, action, science, mathematics[2], and technology. I do not, however, take this as a license to ignore or distance myself from these disciplines and do whatever I want. I hope to show that the approach defended can reveal a coherent picture of both understanding and the understanding subject that nevertheless respects the insights from different fields and addresses its pressing problems (most notably those of epistemology) - but without being constrained by only the focus or insights of those respective fields.

[1] Moreover, the patch of mathematical understanding in particular remains largely unreaped.

[2] The focus here will often include mathematical understanding, being an interesting and peculiar example of something to understand. However, I don't believe there is a conceptual difference between understanding mathematics and understanding anything else, so I shall take the oft-neglected example of mathematical understanding to explicate the epistemic concept of understanding, generally.

DISSERTATION

Overture

Because my approach to understanding does not follow the conceptual or narrative structure of the existing literature (e.g. there are no necessary or sufficient conditions, the notion of beliefs or propositions are not taken as primitive, etc), it will not be helpful to present my own approach as an extension to the existing state of the art. Instead, I will cover the state of the art of the literature only where it is relevant to my own conceptual and narrative structure. I have, however, included a chapter (Chapter 3) that discusses many of the existing problems (and proposed solutions) raised in the literature, and how my approach deals with them, as a means to both contrast and validate my own approach with and over others. Furthermore, I shall refer back and ahead where possible to ensure the reader is on board with what is yet to come or what has been addressed elsewhere.

But in painting a big picture, one must inevitably start somewhere. I have taken pains to divide the big picture into smaller sections (such that we can painlessly focus on details) while also emphasising its relation to the rest of the picture (such that the reader knows how it fits in with the rest). In the beginning, this will involve a lot of internal referencing, and short summaries of future arguments, but the further along we get, the more the pieces will fall into place. Therefore a good place to start would be with a summary of the overall dissertation. The dissertation is divided into two parts. The first part (composed of Chapters 1 to 3) will cover the characterisation of understanding, and the second part (composed of Chapters 4 to 6) with cover the characterisation of the subject with understanding.

The first chapter focuses on characterising the *mark of understanding*. This involves specifying what systematic feature we find so philosophically or epistemically valuable about understanding, and thus find necessary for (and explanatory about) its attribution - regardless of who it is attributed to or what it is about. After considering some proposals (sense-, state- and synonym-based accounts) and their flaws, I shall argue that understanding-attributions always boil down to a particular set of *relevant abilities* (of a subject), composed of acts (salient to the object for a certain context). It will be shown that this is the most coherent and useful conceptualisation of "understanding," because it side-lines the role of feelings (which are a salient, but distrusted aspect of understanding), avoids the pitfalls of mental states (which we can neither discern or value directly) and can more firmly root concepts that do not at first sight seem to match an approach that places its premium on observable acts (i.e. counterfactual acts, beliefs of a subject and the meaning of an object). Having established abilities as the mark of understanding, I will briefly consider some candidate kinds of abilities offered by the

literature as the appropriate one(s) to set up that I will consider none of them as the necessary or sufficient condition(s) for understanding, but instead as what composes understanding. This will allow the quality of understanding to be expressed through the amount of abilities appropriate for a context of attribution - to be further developed in Chapter 2.

In the second chapter, I will conceptualise the *dimensions* and *degrees* of *quality* in understanding, offer up a *contextual approach* to specifying what is salient, and specify some of the problems, opportunities and virtues of evaluating understanding under my approach. I will present four dimensions where a higher degree would lead to a superior understanding. The first is the scope of understanding, which tracks the amount of different abilities. The other three dimensions focus on the degrees of quality within each of these different abilities, and will consist of two parameters (one which widens it and another which deepens it). These dimensions will express how *sensitive* an ability is to the demands of a practice (comprised of the situational responsiveness and accuracy parameters), how *stable* the acts that compose it are across circumstances (comprised of its range and robustness), and how *efficient* the subject is in producing them (comprised of the economy and potential parameters). Unfortunately, and quite unsurprisingly, no single agreed upon universal standard can clarify all attributions of understanding within these dimensions. Therefore, I will conceptualise how to express the contextual variations in each dimension (and each parameter specifically) by allowing the context of attribution to give more or less weight to the salience of specific kinds of abilities, circumstances or efficiencies, along with the option for thresholds. Even if these parameters are imperfect in conceptualising an “ideal” or quantitative assessments of the quality of understanding, they are fruitful in diagnosing the strengths, weaknesses, kinds and differences in quality as well as the problems or opportunities in evaluation (e.g. kludges, indirect vs direct evidence, kinds of understanding complete understanding), as will be shown in Chapter 3.

If abilities are the true mark of understanding, as I will have argued in Chapter 1, then finding counterexamples that showcase we can have understanding without abilities or abilities without understanding would undermine that approach. In the third chapter, I will consider a series of candidate counterexamples and show why each of them fails to hurt the ability approach as presented here. In doing this, I will showcase how my account deals with many of the staple examples to be found in a variety of literatures, and I will further validate my characterisation of understanding as discussed in the previous chapters. It will be argued that most examples that seem to involve understanding without abilities nevertheless bottom out in claims about contextually salient or counterfactual abilities. Next, it will be argued that examples that seem to involve abilities without

understanding are nevertheless ultimately justified through a lack of abilities (conceptualized through the parameters of Chapter 2), or due to focusing on the wrong subject. This latter option prepares us for the last three chapters, where I will consider the subject with understanding, and whether this can ever apply to an extended entity, a group or an artificial system.

Understanding is always predicated on a subject. Therefore, the fourth chapter is focused on the *mark of epistemic subjecthood*, a focus which is sorely lacking from the epistemology literature (especially regarding understanding). Is there a systematic way to reveal what is required for subjecthood before we can attribute it with epistemic properties (such as understanding)? Answering such a question involves a conceptualisation of what makes up a subject and what doesn't, drawing a line between what lies within the boundaries of the subject and what doesn't, as well as why we draw that line where we do. While the answer may change depending on what one is interested in, I will argue that a good guideline is to let the boundaries be dictated by what implements a coherent and persisting epistemic agent. This means that, as a mark of epistemic subjecthood, I will defend the *interpretationist* approach, and more particularly the *epistemic stance* (the intentional stance with an epistemic focus). The epistemic stance is the instrumental strategy of interpreting behaviour by treating it as if the entity were governed by beliefs, epistemic aims and epistemic tactics (as well as any other intentions that play a supporting role). Having defended the epistemic stance as the mark of epistemic subjecthood, I will argue that if an entity, composed of more than just a human individual, can grant us explanatory or predictive powers through the epistemic stance, then taking advantage of this power is not only warranted and fruitful, but consistent with our best conceptualisations of individuals. In that case, we are dealing with an extended epistemic agent. To end, I will discuss 7 different cases to showcase what can get "extended" in *extended understanders* and how. This includes socially extended understanding, on which I will elaborate in Chapter 5.

In the fifth chapter, I turn my focus to the notion of *collective understanding*. There are countless examples in natural language where groups are attributed with understanding. Are these attributions supposed to be merely empty rhetoric, superfluous metaphors, and convenient shorthands, or is there any genuine explanatory power to them? While I will not conclusively answer whether we can find any existing group epistemic agents, I will shed light on the conceptual space involved in substantiating such an answer. I will argue that a couple of basic steps need to be traversed for a group to warrant the attribute of genuine collective understanding. First and foremost, there needs to be a group, along with whatever that entails. Secondly, that group needs to, as a group, display abilities (because there is no collective understanding without the trait of understanding). And thirdly, those abilities need to

result in a successful epistemic stance (because there is no collective understanding without an epistemic subject to attribute it to). However, even if a group of human individuals forms a body that acts as one (thus creating an explanatorily powerful target of the epistemic stance), it may yet be possible to reduce that group-level explanation to individual-level explanations, making the appeal to a collective subject superfluous. This is the *reducibility problem*. When is such reducibility a problem and when is it not? I shall argue that reducibility is a problem when the abilities and epistemic agency of the group can be straightforwardly mapped onto a conglomerate of those of its members (because there is no collective understanding if the attribution is not uniquely tied to the group). So the last step will involve pointing to the lack of such a mapping relation due to emergence. To end, I will give both an idealised example as well as a brief indication of real world examples.

In the sixth and last chapter, I will consider the notion of *artificial understanding*. Can artificial systems (such as computers) ever be attributed with understanding? To answer this question in the positive will involve first establishing whether it is possible for artificial systems to display epistemic abilities, and whether such abilities allow the epistemic stance to be explanatory or predictive. There are, however, some criticisms that take artificial understanding to be impossible in principle, even in spite of the presence of abilities or the success of the epistemic stance. They involve the regress, reducibility and rigidity problem and they form three conceptual hurdles that we will have to overcome to justify the conceptual possibility of artificial understanding. Overcoming the first two conceptual hurdles involves having an answer to the question: Why doesn't an artificial system's purported abilities or epistemic agency (and therefore understanding) automatically *regress* to its programmer or *reduce* to its programming? I will admit that if you can straightforwardly map the abilities or epistemic properties of the artificial system to those of its programmer or to the procedures in its programming, it will be explanatorily superfluous (even if convenient) to postulate an additional agent and attribute the abilities to the system as a whole. But the regress and reducibility problem take the legitimate worry of an explanatorily superfluous epistemic stance and unduly extend it to any case where there is a causal origin or supervenience, no matter how convenient, self-sufficient or distinctly explanatory it is to consider the entity by itself. Overcoming the third conceptual hurdle involves being able to answer why artificial systems aren't too *rigid* to display the full scope and sensitivity of abilities we find in human beings. I will argue that the rigidity problem mistakenly assumes that the level of computation must align with the level of abilities, when the computational level may fall well below that level. Having addressed the three conceptual hurdles, I will end this chapter by giving examples of how the road to artificial mathematicians is being trodden in the wild.

PART I

CHARACTERISING UNDERSTANDING

PRELUDE 1

That Within Which Passeth Show

In the University of Wittenberg, a group of students are waiting in line for the feedback of their mathematics exam given by Professor Gertrude Ryle. A student comes out of her office.

ROSENCRANTZ: How did it go?
GUILDENSTERN: Not very well at all, I'm afraid. I thought I had it, but apparently I got everything wrong. How did it go with you?
ROSENCRANTZ: Well, I was very nervous going in, but it actually seemed to go alright.

After a little while, another student comes out.

GUILDENSTERN: So, Hamlet, how did it go with you?
HAMLET: It went very well. The professor ended by saying "You seem to understand the material."
ROSENCRANTZ: That's reassuring!
HAMLET: To which I said:
"Seems," madam? Nay, I do. I know not "seems."
'Tis not alone my derivations, professor,
Nor customary rules of inference,
Nor windy explanation of proof outlines,
No, nor the fruitful heuristics used,
Nor the representations or generalisations,
Together with all forms, kinds, shapes of acts,
That can denote me truly. These indeed "seem,"
For they are actions that a man might play.
But I have that within which passeth show,
These but the trappings and the suits of understanding.
ROSENCRANTZ: Did you really say that?
HAMLET: No, but I should have. Anyway, this was my last exam, so I'm going home to Elsinore.

Hamlet leaves.

GUILDENSTERN: Oh. That's actually very reassuring!
ROSENCRANTZ: Oh. That's actually very worrying!

Spurred on by Hamlet's words, Rosencrantz & Guildenstern both decide to go back into the Professor's Office to raise a complaint. Rosencrantz takes the lead.

ROSENCRANTZ: I have a complaint.
PROFESSOR: Oh?

ROSENCRANTZ: You've given me a passing grade, praising me for my excellent understanding of this proof above all else. But the truth is, I don't deserve the credit. I don't understand it at all.

PROFESSOR: Did you copy your neighbour's answers?

ROSENCRANTZ: No, of course not. I would never.

PROFESSOR: Oh dear, did my exam leak online and did you memorise the answers?

ROSENCRANTZ: No, I didn't cheat. I worked out the proof on the basis of the axioms given. I actually find that easier than trying to memorise a whole proof verbatim.

PROFESSOR: That's normal, I have the same. It's easier to just see the logic of it.

ROSENCRANTZ: That's exactly the problem. I don't *see* it.

PROFESSOR: Then how come you gave such an appropriate answer? You gave every indication that you did see it.

ROSENCRANTZ: I have the suit of understanding, but not its features. I can easily conjure up the proof, I can tell you what would happen if the axioms were otherwise, I can reshape its physical representation from text to a visual diagram, and all that -

PROFESSOR: Ah, do you mean that these abilities seemingly come out of nowhere, outside of your control?

ROSENCRANTZ: Oh no, sorry to confuse you, Professor. No, I'm in perfect control of my abilities. If I'm struggling with a problem, I always deliberately use mathematics to help me solve it. And it's always worked out great.

PROFESSOR: Okay... So do you mean you can't explain what it is that you're doing? The solutions just present themselves?

ROSENCRANTZ: Quite the opposite, professor. I can give you every skipped lemma, every precise step in my methodology, whether it is rigid and formal inferences, or fluid and fuzzy associations.

PROFESSOR: So what's the problem?

ROSENCRANTZ: It's just that I don't see it. See what you see, I mean. I don't have that within which passeth show. I strongly suspect that my abilities actually come from a completely distorted mental representation of mathematics.

PROFESSOR: But you always describe your vision so vividly and with such detail. And all of it is exactly appropriate.

ROSENCRANTZ: Sure, I can *describe* it appropriately, but I still suspect that what I *see* is actually entirely inappropriate.

PROFESSOR: On what basis can you suspect it inappropriate when you have absolutely every indication to show they are appropriate?

Guildenstern, starting to get impatient, now interjects at this point.

GUILDENSTERN: Might I interject at this point? I too have a complaint. You've given me a failing grade, stating that I don't, for instance, understand this proof at all. But I do.

PROFESSOR: You haven't given me any indications that you have. All of your answers on the exam were incorrect - not even close to correct, in fact! You don't seem to see what the proof is about at all.

GUILDENSTERN: That's exactly where you're wrong, you see. I see it all. *But* I can't put it into words. It's a problem I have. I have the features of understanding, but not its suit.

PROFESSOR: Which features are those then?

GUILDENSTERN: Well, I have, in my mind's eye, a vision of mathematics in its entirety. Anything you can think of, I have a representation of it. For instance, I'm sure you also see why the square root of 2 must be rational?

PROFESSOR: You mean irrational.

GUILDENSTERN: Sure, irrational. Well, do you?

PROFESSOR: Yes, I can think of the proof.

GUILDENSTERN: Do you see why it has to be irrational?

PROFESSOR: Yes.

GUILDENSTERN: Me too! I see the exact same thing! The proof is in my mind. Clear as day.

PROFESSOR: How about you describe what it is you're seeing?

GUILDENSTERN: I'm afraid that won't work, it's ineffable.

PROFESSOR: Is it visual?

GUILDENSTERN: Not always, sometimes I see the proof in my mind's eye, sometimes I feel it with my mind's hands or taste it on my mind's lips. But there is always a mind's sense that picks it up.

PROFESSOR: Well, I'm afraid "seeing" is merely metaphorical. Understanding mathematics isn't being a passive witness to representations of mathematics. My little 4 year old daughter looks at my work all the time, but she doesn't understand what she's seeing.

GUILDENSTERN: Why not?

PROFESSOR: She doesn't *grasp* what she's seeing. She sees it, but she just can't… manipulate it.

GUILDENSTERN: I see. You mean to say that the difference between her physical seeing and your metaphorical seeing is that you can mentally manipulate what you're seeing?

PROFESSOR: Yes, I suppose so.

GUILDENSTERN: So can I! My mind's hands are constantly grasping at mathematics, changing postulates and affecting the axioms.

PROFESSOR: You mean the other way around?

GUILDENSTERN: Sure. Sorry about the mistakes, I have to remind you that I can't describe what it is that I'm seeing. I'm doomed to always describe it entirely inappropriately even though what I see and grasp is always completely appropriate.

PROFESSOR: You sure don't act like it.

GUILDENSTERN: No, no, no. You've got it all wrong. You can't *act* understanding. The quality of understanding is nothing to do with witnessing someone act - It's not deriving axioms, explaining a proof-outline or representing the proof - that isn't what makes it understanding. It's just a person having the right mental state.

PROFESSOR: On what basis can you think that what it is that you're supposedly seeing, feeling and tasting is appropriate when you have absolutely no indication to show for it?

Chapter 1

THE MARK OF UNDERSTANDING

Understanding is a valued trait in any discipline and it is fair to say that we sometimes uncontroversially attribute it to certain subjects. But is there a systematic way to reveal what understanding-attributions are or should entail? What is it in those situations of uncontroversial attribution (or dismissal) that guides us - or should guide us - in ascribing (or denying) understanding? Answering this question involves specifying what systematic feature we find so philosophically or epistemically valuable about understanding, and thus find necessary for (and explanatory about) its attribution - regardless of who it is attributed to or what it is about.[3] I call this the "mark of understanding"[4], because it is what *demarcates* it. Characterising this mark is what the focus of this first chapter will be.

After considering some proposals (sense-, state- and synonym-based accounts) and their flaws, I shall argue that understanding-attributions always boil down to a particular set of relevant abilities (of a subject), composed of acts (salient to the object for a certain context). It will be shown that this is the most coherent and useful conceptualisation of "understanding", because it sidelines the role of feelings (which are so distrusted) and avoids the pitfalls of mental states (which we can neither discern or value directly).

I will briefly consider some candidate kinds (or brands) of abilities offered by the literature to be the appropriate one(s), to set up that I will consider none of them as the necessary or sufficient condition(s) for understanding, but instead as what *composes* understanding. This will allow the quality of understanding to be expressed through the amount of salient abilities (which will be further developed in Chapter 2) and will allow understanding to vary the salience of these abilities along with the meaning of the object that is being understood (which we can conceptualise as the appropriate usages or indications thereof) within a context of attribution (which will also be further developed in Chapter 2).

[3] Nonetheless, whatever marks understanding needs to be able to be applied consistently to various human subjects (and possibly beyond - which will be the focus of Chapters 4 through 6) and across various objects with varying degrees of (contextual) quality (which will be the focus of Chapter 2). Furthermore, it needs to allow us to deal with the known philosophical problems of marks, as well as address possible counter-examples (which will be the focus of Chapter 3).

[4] The term is in the same spirit as the "mark of the cognitive" (Adams & Aizawa, 2001, p. 46) as used in philosophy of mind.

To end, I will cover some of the useful concepts associated with understanding that do not obviously match an act-based approach and show how they can not only keep their explanatory power, but are even more firmly rooted as instrumental concepts derived from acts. This includes the *modality of abilities* (i.e. would have displayed the ability if…) as the explanatory generalisation of conditional acts, the *meaning of objects* as indications of their appropriate usage (according to a particular practice - to be further developed in Chapter 2) and the *mind of a subject* (i.e. its beliefs, aims, etc) as an explanatory interpretation of the subject's behavioural profile (to be further developed in Chapter 4).

1.1 The Value of Understanding & Its Mark

In this chapter, I shall argue that understanding-attributions always boil down to the subject in question possessing a particular set of relevant abilities, and that this is the most coherent and useful conceptualisation of the notion of "understanding". But before I propose a conceptualisation of understanding, it will be worthwhile to consider what makes both understanding and its mark valuable in the first place.

The Value of Understanding

Humans strive for understanding. This is quite a bold and yet uncontroversial claim. Understanding is a predicate applicable in just about any activity or discipline, and is considered a valued trait or aim for anyone involved in them. This is as true for physics (e.g. understanding the implications of relativity theory, failing to understand quantum mechanics), politics (e.g. understanding the root of the problem in an international conflict, understanding the difference between anarcho-communism and state communism), baking (e.g. understanding why the pie has a soggy bottom), psychology (e.g. understanding why someone is depressed), languages (e.g. understanding Chinese), literature (e.g. understanding why a text resonates with an audience or which context contributes to that effect), biology (e.g. understanding why DNA has a double-helix shape), medicine (e.g. understanding why the medicine is more effective through oral administration) or mathematics (e.g. understanding why the denial of a theorem would result in contradictions). Both "scientists and laypeople alike will typically regard understanding as one of the most important and highly valued products of scientific research and teaching." (de Regt, 2017, p. 1) The promise of (a better) understanding is a drive behind the sciences, art or religion. (Baumberger et al, 2016)

Even in mathematics, often taken to be a special case within the sciences, understanding takes a central role. That students should *understand* mathematics is "[o]ne of the most widely accepted ideas with the mathematics education community" (Hiebert & Carpenter, 1992, p. 65) and achieving this is

"[t]he main goal of elaborating teaching designs, projects, new software and textbooks." (Sierpinska, 1990, p. 24) Yet understanding is not confined to education alone. To gain understanding is a key motivation even for research mathematicians. This was made all the clearer when computers started contributing to their field. When the Four Color Theorem was proved by a huge amount of automated testing (Swart, 1980), one of the most heavily discussed aspects was the value of this contribution. For instance, Frank Bonsall, a leading mathematician, said:

> "We cannot possibly achieve what I regard as the essential element of a proof —our own personal understanding— if part of the argument is hidden away in a box." (Bonsal, 1982 quoted in MacKenzie, 2004, p. 102)

When it comes to mathematics, understanding is not only a motivator to prove (Rav, 1999), but also a motivator to re-prove what is already known, but not understood (Dawson, 2006; Thurston, 1998), and understanding is a possible criterion to distinguish proofs that merely demonstrate from proofs that explain (Delarivière, et al, 2017; Frans, 2020). It may even be that understanding is the driving force of mathematical practice a lot more than formal validity is.[5] (Delarivière & Van Kerkhove, 2017). When someone studies or practices mathematics, their aim is not usually to only acquire or contribute to a list of mathematical results, but also to *understand* them.

The value of understanding has been explicitly discussed in the field of epistemology (see Grimm, 2012 for an overview), with some authors (e.g. Zagzebski, 2001; Kvanvig, 2003; Pritchard, 2009) even arguing that its value exceeds that of knowledge, because "we would surely rather understand than merely know'' (Pritchard, 2009, p. 30). This distinction in value does not, however, enjoy a consensus, with disagreement resting largely on the proposed distinctions[6] between the nature of understanding and knowledge[7]. But whatever the relative value between knowledge and understanding, or the distinction that motivates it, it is not the value of understanding that is at issue.

Some (e.g. Griffin, 1984) go as far as saying that the value of understanding is an intrinsic one. It is certainly possible to dispute that claim, but what would be harder to dispute is that understanding

[5] We'll come back to this notion in Section 6.3.

[6] The proposed difference-maker ranges from understanding's transparency (Zagzebski, 2001) or coherence (Kvanvig, 2003) its status as an intellectual achievement or cognitive ability (Pritchard, 2009). Each of these will be discussed in this dissertation where relevant (transparency in Section 1.3, coherence in Section 2.1 and 3.3, and intellectual achievements or cognitive abilities in Sections 1.2 and 3.2).

[7] If understanding is simply a type of knowledge, then the issue is a non-starter based on an unfair comparison. (Brogaard, 2005).

nevertheless brings with it certain valauble benefits (e.g. being able to predict an outcome, improve powers, avoid dangers,...). Where there is understanding, there is an opportunity to exploit that understanding for your own gains. Nowhere is this idea given more force than in Marie Sklodowska-Curie's famous quote:

> "Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less." (quoted in[8] "Marie Curie," 2020)

The Value of a Mark of Understanding

If understanding is a valuable epistemological trait, then we need a fruitful epistemological characterization of it. Understanding is itself yet another target which we may strive to understand. And yet, the notion of understanding has, in the past, largely resisted characterisation from philosophers, who have instead tended to regard the concept with much suspicion, or not much at all.

In the field of epistemology, the related concept knowledge - which has enjoyed a long tradition of philosophical investigation - has received the characterisation of "justified true belief"[9], along with explications on what that could or should entail, as well as an open discussion about any further conditions or constraints, sparked by Gettier (1963). But understanding doesn't share this longevity in investigation or richness in characterisation. This may, in part, be due to a simple lack of differentiation in terms. Locke's *An Essay Concerning Human Understanding*, for instance, used "understanding" as a synonym for "knowledge". Both derive from the word "episteme", but the focus in philosophy has evolved, or shifted (Grimm, 2012), into a characterisation of what we today call "knowledge" and not what we call "understanding". (Baumberger et al, 2016) This to the lament of recent epistemologists (e.g. Zagzebski, 2001; Kvanvig, 2003; Elgin, 2007; Pritchard, 2009). While claims of knowledge can be corroborated or contested with philosophical criteria, when it comes to understanding – which we established to be of no less value - we are reduced to intuition and hand waving.

If understanding is a central aim of many of the sciences, as we saw earlier, would it not be useful to know what the target is, so that they may improve their aim? Yet even in the field of philosophy of science, the concept of understanding was long left out of focus in favour of other concepts. Understanding has close ties with the concept of explanation, which has benefitted from a lot more attention in its literature. The tie is not hidden. In fact, "virtually every theory of explanation also

[8] Sadly, I could not find the primary source for her quote.

[9] A characterisation which is not quite as old as it is deemed to be. See (Dutant, 2015)

places a premium on the power of an explanation to produce understanding" (Trout, 2005, p. 198).[10] Yet this did not curry attention to the concept of understanding in philosophy of science. The justification for this lack of attention was understanding's "pragmatic" nature. If we focus on understanding, so the argument goes, we are limited to investigating the subjective and relative responses of a single individual.

> "Very broadly speaking, to explain something to a person is to make it plain and intelligible to [her], to make [her] understand it. Thus construed, the word 'explanation' and its cognates are *pragmatic* terms: *their use requires reference to the persons involved in the process of explaining*. In a pragmatic context we might say, for example, that a given account A explains fact X to person P1. We will then have to bear in mind that the same account may well not constitute an explanation of X for another person P2, who might not even regard X as requiring an explanation, or who might find the account A unintelligible or unilluminating, or irrelevant to what puzzles [her] about X. Explanation in this pragmatic sense is thus a relative notion: something can be significantly said to constitute an explanation in this sense only for this or that individual." (Hempel, 1965, p. 425-426, italics added)

It is true that concepts like understanding are sterile without a reference to a subject. For instance, it is *Marie* who understands why the global temperatures are rising, *Bill* who understands why the square root of two must be irrational, and *Wendy* who understands why the old theory failed to account for the facts. In this pragmatic context, whether an explanation indeed provides understanding is dependent not only on the explanation, but on the agent receiving the explanation, her beliefs at the time, her intelligence, her critical standards, her personal idiosyncrasies, and so forth. This, according to Hempel (1965), is the wrong focus. The pragmatic concept may be of interest to psychologists or educators, but not to epistemologists or philosophers of science, where the focus should *not be* on whether an argument is subjectively effective for a particular individual, but whether it objectively explains *the facts*. There is a credit to this evasion of the pragmatic, in the sense that the concept of explanation is intended to cover more than its effects on a particular individual. But there is also a danger to it, in the sense that the concept of explanation would be vacuous if it is independent of any and *all* individuals. Remember the premium that accounts of explanation place on its power to impart understanding. If a particular candidate argument provides no understanding to *any* individual

[10] This makes the concept of explanation implicitly tied to the concept of understanding. Though not necessarily vice versa. Lipton (2009) argues that one can acquire understanding without an explanation. If this is true, then an account of understanding is less dependent on an implicit account of explanation than the other way around.

in humanity, but it does fit a philosopher's account of explanation, do we fault humanity or the philosopher? Surely it is the latter.

So even within the philosophy of explanation, the concept of understanding is not sterile. Furthermore, the fact that it is pragmatic does not entail a lack of objectivity (unless in a very narrow sense of the word). So much has been argued for by, for instance, Friedman (1974) and further developed by de Regt (2017). In the subsequent chapter, I will show how, under my account, there is a place for scientific standards in the evaluation of understanding. So the pragmatic nature of understanding is no justification for its conceptual exclusion, and its conceptual exclusion may lead to a more sterile approach to related concepts.[11]

So far, we've only focused on the value of the mark of understanding as it applies to humans, but the concept of understanding is equally beneficial to *android epistemology*, where the aim is to have a better grasp of the process and limits of knowledge and understanding in artificial agents. (see e.g. Ford et al, 1995; Ford et al, 2006; Delarivière & Van Kerkhove, 2017) The use of computers in mathematics research provides yet another interesting example here, because it provoked a fundamental discussion as to their epistemic capacities. The discussion centered on three issues: (a) reliability, (b) surveyability or intelligibility and (c) its providing or being provided by understanding. Based on one or several of these, some people considered computer proofs to be: uninteresting or unsatisfying mathematics, a completely different sort of mathematics, or no mathematics at all. (MacKenzie, 1999; Vervloesem, 2007) The issue sparked a debate about the differences or similarities between computer proof and its traditional human counterpart. However, both computers and humans are subject to reliability and (sometimes) surveyability issues, making it hard to argue for a dichotomy between the two. (Delarivière & Van Kerkhove, 2017) Nonetheless, humans are considered as more trustworthy due to another quality they possess or supply. For instance, the mathematics community accepts peer-reviewed results without everyone partaking in this process, instead allowing peer reviewers to function as the testimony of trustworthy black boxes (Geist et al. 2010). What human surveyors (in the best cases) supply to warrant peer review and what provers supply that empower their proofs is (c) understanding. According to Rav (1999), this focus on understanding means the primary goal of mathematics is the development of mathematical meaning which cannot be derived from *formal* expressions, but instead requires active interpretation, an "irreducible semantic component" (Rav, 1999, p. 11). On the basis of such claims, computers get pushed outside

[11] See (Delarivière, Frans & Van Kerkhove, 2017) for elaborations on how the exclusion of the role of understanding restricts accounts of explanation.

the realm of understanding and thus the locus of mathematics. While this lack of understanding often gets mentioned (MacKenzie, 1999) and is assumed to constitute a necessary difference or dichotomy, the critique is vague and little is done to explicate or investigate what this *informal* understanding might actually or preferably entail as well as when any of its characterizing criteria are met or left unsatisfied. Avigad (2008) laments this lacuna in philosophy of mathematics in particular:

> "[T]here is a smaller, but significant, community trying to automate mathematical discovery and concept formation (...) If there is any domain of scientific inquiry for which one might expect the philosophy of mathematics to play a supporting role, this is it. The fact that the philosophy of mathematics provides virtually no practical guidance in the appropriate use of common epistemic terms may lead some to wonder what, exactly, philosophers are doing to earn their keep." (Avigad, 2008, p. 313-314)

Only if we have a better understanding of the concept of understanding, are we in a position to evaluate the success and alleviate the failures (or stipulate its in principle impossibility) of understanding in artificial systems. The same can be said for disciplines other than mathematics and the same holds true for other unconventional subjects such as extended systems or groups. That is why the second half of this dissertation will utilise what has been set up in the first half to deal with such unconventional subjects, namely extended systems (Chapter 4), groups (Chapter 5) and artificial systems (Chapter 6).

Not only is understanding's pragmatic nature no justification for evading its mark (which could even prove detrimental to concepts that are directly or indirectly tied to it), the pragmatic nature of understanding also makes clear that the potential and flaws of the targeted subjects cannot be adequately expressed without a mark of understanding. So if understanding is a valued trait and aim comparable to that of knowledge, its mark deserves to be described with philosophical care equal to that of knowledge.

The Mark Question

Do we really need a complex and coherent philosophical characterisation of understanding before we may talk of its applicability? If we ascribed value to understanding, then it should follow that we have at least *a sense* of what the target of that value-ascription is. It would be hard to value something without a sense of *what* it is that is being valued. Of course, this doesn't entail that we can give a full conceptual description of it, but we may be able to "recognise it when we see it". This may very well

be true. Nonetheless, it would be easier to find if we know where to look, and we would see it more clearly if we know where it begins and ends. In other words, both the pursuit and demarcation of understanding would be facilitated by having a clear and coherent philosophical concept that marks it out. So, when we attribute someone with understanding, is there a systematic way to reveal what that should mean? This is a question where philosophy can be of help. A growing movement in both philosophy of science and epistemology has granted that understanding is a topic that merits philosophical discussion. What could or should constitute genuine understanding has not, however, enjoyed large agreement - as we shall see in the subsequent sections.

Where do we begin in constructing an answer? Perhaps by demarcating what should take precedence in the discussion. To do that, allow me to formulate a simple claim about what we are talking about when we talk of understanding, namely that:

> Understanding is a trait (T), of a subject (S), concerning an object (X)

I've proposed this fairly uncontroversial sentence such that, if we unpack it, three aspects of understanding present themselves for characterisation, namely (T) the trait itself, (X) the object it concerns and (S) the subject that possesses the understanding. This allows us to branch the topic of understanding in three different sub-topics, each focusing on a different aspect and each of which can be studied and characterized semi-independently of one another. I say "semi" because the intention is not to sever the connections between the branches, but merely to prevent their conflation. Each of these branches will be addressed in their own turn. Most epistemological treatments of understanding seem to focus on only one or two of these (mostly the property or object), while leaving the others ignored or implicit. I'll hope to show that some of the convoluted problems this creates can be easily avoided by taking a step back and seeing the full coherent picture. This is one of the reasons why I will present a big picture approach to understanding.

The first conceptual branch is that of the *trait of understanding* (T). Before we can begin spelling out who it is that understands and what it is that they understand, as well as why this is so, we need to be able to specify what it is that makes us ascribe understanding, regardless of who it applies to or what it is about. Obviously, this is not to deny the pragmatic nature of understanding by treating understanding as if it were some ontological entity detached from any individual. This is just to focus our attention on what it is that these ascriptions of understanding (to various subjects regarding various objects) have in common. What is it that *marks* the understanding? It is important that we

start with the mark, and not its corollaries, so as to weed out the possibility of conceptual entanglement. As a conceptual branch, it has been the primary focus of the epistemology literature concerning understanding, and it will be the main focus of this first chapter. Therefore, the mark defended in this chapter will need to allow us to deal with the known philosophical problems of marks, as well as address possible counter-examples (which will be the focus of Chapter 3).

The *object of understanding* (X) has, as a conceptual branch, received the most attention of all, namely in the philosophy of education literature. Most articles that deal with educating students on a particular topic and how to do it well, will focus on how to help students reach understanding of a specific object, X, and how to evaluate what they have learnt. While this literature is interesting, their goal is practical and local, so is not usually focused on making explicit the notion of understanding as a whole[12] or how to conceive of the different kinds, ways and qualities (as well as their contextual variations) in which the mark manifests itself. In Chapter 2, I will elaborate on this specifically, but for our current purposes, it is important to keep in mind that the mark of understanding needs to allow manifesting itself differently depending on the various objects of understanding, and indicate a way to conceptualise these differences.

This leaves us with the *subject with understanding* (S). As a conceptual branch, it has been severely under-nurtured in the philosophy of science and epistemology literature. With the exception of Toon (2015), I cannot cite a single article that explicitly covers what it takes for understanding to belong to a subject and what makes out that subject. This is quite strange for a concept which is pragmatic (i.e. predicated of a subject). That is not to say, however, that there is no literature to draw from. The subject with *knowledge*, for instance, is a topic which has received some substantial attention in recent years, by drawing from other lucrative fields, such as philosophy of mind, and cognitive science. The entire second half of this dissertation (i.e. Chapters 4 to 6) will be spent to consider the subject with understanding with similar care, building on what has been set up in the first half, and drawing from the relevant aforementioned literatures. But for now, it will be important to keep in mind that whatever marks understanding needs to allow manifesting itself in various subjects.

1.2 Dispelling Sense, State & Synonym Accounts

So let's consider some proposals of what it is that a subject could (or should) possess when we attribute her with understanding. This involves specifying what we find so philosophically or epistemically valuable about understanding and thus necessary for its attribution, regardless of who

[12] There are exceptions, such as (Sierpinska, 1990; 1994).

possesses it or what it is about. For instance, something which is invariably present in an account of understanding, be it as a symptom or mark, is the presence of *abilities*. I will soon argue, with inspiration drawn from Ryle (1949/2000), why treating abilities as a mark rather than a symptom is more philosophically sound. But first, I will first consider some candidate marks that don't put abilities front and center. These can have some intuitive and philosophical plausibility, but also lead to troubles on both those fronts, indicating that they are neither as plausible nor as intuitive as one might first assume.

Sense

Certainly the most salient feature of understanding is the *feeling* or *sense* of understanding. The most notable of feelings related to understanding is the aha-erlebnis.[13] But we can also feel confident in our newly-found epistemological powers, take pleasure in finding transparency or coherence where we thought there was opaqueness and disjoint, and we can take satisfaction from fulfilling a drive to find evidence that support our theories or modify our theories to fit the evidence (Gopnik, 1998). The sense or senses we associate with understanding can be spread out in time or hit us with a flash, a moment of eureka. This latter sense is so common to us that it sounds familiar to scientists and laypeople alike.

But appearances can be deceptive, and this sense has rightly been criticised (e.g. Trout, 2002) as an unsatisfying characterisation of understanding as an epistemological mark. It is not difficult to think of examples where someone has genuine understanding without having any accompanying feeling and even easier to think of examples where someone genuinely experienced that sense, but was simply mistaken. Trout (2002) attributes this to a combination of hindsight bias (the observed effect of people systematically overestimating their past powers) and overconfidence bias (the observed effect of people systematically believing they are right even when they aren't). Indeed, the feeling of understanding may be its most familiar and salient aspect, but it is almost unanimously agreed (e.g. Kvanvig, 2003; de Regt & Dieks, 2005; Wilkenfeld, 2013b; Toon, 2015; Ylikoski, 2009) that it is neither necessary nor sufficient for genuine understanding. For this reason no one defends it as the *mark of understanding* (unless to discredit understanding as a concept) and detaching understanding from its associated sense has been the start of taking the concept of understanding seriously.

This doesn't entail that the role of feelings are epistemologically irrelevant, a mere epiphenomenal by-product that is no more than the "irrelevant phenomenal steam that the brain lets off in the course

[13] The aha-erlebnis is not easy to reduce to other feelings, like surprise, familiarity or expectation (see Lipton, 2009).

of its serious cognitive work" (Lipton, 2009, p. 55). Feelings may bear some epistemological relation to understanding other than as a mark. For instance, the sense could serve as indirect evidence or a strong heuristic in the pursuit of understanding. For instance, under the condition that one's background beliefs are varied and true, the satisfaction of new information cohering with these background beliefs makes the sense of understanding to be a reliable indicator of understanding. (Grimm, 2011) Given that this feeling seems generally desirable, it may also illuminate, at least in part, why we value and pursue understanding. Indeed the psychological payoffs are more local and more immediate in the pursuit of understanding than some of the epistemological benefits. (Lipton, 2009) Gopnik (1998) argues that this sense of fulfillment is due to a useful evolutionary drive for theory-formation (in a similar way that the orgasm is due to a drive for baby-making).[14] So even if a sense is neither necessary nor sufficient, its presence may significantly correlate with acquiring a better understanding.

But the role of feelings could be more epistemically embedded still. Historically, people have been naive about what sort of qualities or processes underlie human intelligence, so it is conceivable we are underestimating the role of feelings in understanding. It is entirely possible that an agent that lacks such a psychological dimension is condemned to lack in understanding also. This would disprove de Regt's notion that that a feeling of understanding "has no epistemic function." (de Regt, 2009, p. 587)[15] But this is mere speculation. Others have pointed that feelings can also play an epistemically misleading or inhibiting role. For instance, feelings can be misleading due to overconfidence and hindsight bias (Trout, 2002) or play a predominantly inhibiting role due to overestimating one's detail, coherence, or depth of understanding (Ylikoski, 2009). To the extent that a psychological sense of understanding plays an epistemic role, it will merit inquiry in the context of epistemology, but there is no consensus on whether it does. What does enjoy widespread agreement is that it is not the *mark* of understanding.

States

Philosophical conceptions of understanding invariably invoke the presence of mental states (be it as explanatory concept or trait) and abilities (be it as trait or symptom). So two main lines of spelling out the trait of understanding are as either (a) appropriate mental states, potentially with abilities as

[14] "From our phenomenological point of view, it may seem to us that we construct and use theories in order to achieve explanation or have sex in order to achieve orgasm. From an evolutionary point of view, however, the relation is reversed, we experience orgasms and explanations to ensure that we make babies and theories." (Gopnik, 1998, p. 102)

[15] de Regt (2004) concedes that it can be "a source of motivation" (p. 104), but not an aim of science.

symptoms (e.g. Zagzebski, 2001; Grimm, 2011, 2016; Van Camp, 2013; Wilkenfeld, 2013b) or (b) appropriate abilities, potentially with states as explanatory constructs (e.g. de Regt & Dieks, 2005; Ylikoski, 2009; Hills, 2015; Delarivière & Van Kerkhove, 2017). Siding with the latter, I will argue why considering abilities as a mere symptom of mental states would put the cart in front of the horse, creating a variety of needless problems. The argument in short is that this is because it is not mental states themselves that are empirically accessible or epistemically valuable to us. We both detect and judge mental states by the abilities, and not vice versa. Allow me to elaborate.

To start, it is worth differentiating a mental state from a physical state such as a brain state. To my knowledge, no one defends physical states as the mark of understanding. It is quite obvious that those physical states themselves are neither what we value about understanding (e.g. we do not say "ultimately, it is the aim in science for our brains to reach state X") nor what we look for in evaluating someone's understanding (e.g. we do not say "Did you understand what I said? Let's get you in the brainscan and find out"). If we characterise understanding through *mental* constitution or occurrences (conscious or subconscious), then the successes of an understanding subject will need to come from the subject having something appropriate (such as mental representations) in her metaphorical mind's eye or other mind-organ, and the abilities that the subject may display will merely be the fortuitous public effects of those appropriate private occurrences. We can find some explicit or tacit endorsement of this view in characterisations of understanding (e.g. Hiebert & Carpenter, 1992; Zagzebski, 2001; Barmby et al, 2007; Wilkenfeld, 2013a and arguably Grimm, 2011, 2014). For instance, Wilkenfeld (2013) characterises understanding as such:

> "A statement, attributed in context C, that thinker T understands object o, is true if and only if T possesses a mental representation R of o that T could (in counterfactuals salient in C) modify in small ways to produce R', where R' is a representation of o and possession of R' enables efficacious (according to standards relevant in C) inferences pertaining to, or manipulations, of o." (Wilkenfeld, 2013a, p. 1003-1004)

The mental representation here functions as the mark of understanding. This view fits with some of our everyday language about understanding and its justifications (e.g. seeing how things fit together, pointing to gaps or flaws in a mental image, or abilities being treated as evidence rather than proof of understanding). I will argue, however, that marking understanding via mental states leads to problems both practical and metaphysical.

Firstly, if understanding is characterised through the appropriate mental occurrences, then the task of gauging another person's understanding is not just difficult, it is impossible. We can't empirically discern what people see "in their mind's eye" or what "state their mind is in," except by inferring this through what they do (e.g. through what they say) - in which case the causal role of their internal imagery or mental state is an instrumental postulate. We cannot look into anyone's mind, private as it is. We would have no idea whether Euclid actually understood geometry, for only Euclid himself knew what went on in the privacy of his own mind. So, either our attributions of understanding are speculative exercises in psychology or attributions of understanding are not really about these private occurrences. The only mental representations we have access to is our own, and even then it depends on what we mean with "access" and "representation". Of course, Euclid can *tell* us about what goes on in his mind, but saying is a kind of doing and if his private occurrences are what mark his understanding (more than his doings or sayings), then we have absolutely no way to check whether he isn't subtly or unequivocally deceiving us. In fact, we don't even have any guarantee that he isn't mistaken himself. We'll return to this possibility soon. What subjects with understanding share (or overlap in) is their abilities, not their presumed states.

Maybe there is a way out of this conundrum by reinvoking physical states. Most philosophers and scientists agree (me included) that there is *some* relation between what the brain is doing and what the mind is doing (more on that in Part II). If we would want to be able to take advantage of that relation to detect or mark understanding, our bets would be safest with the *identity theory of mind*. According to the identity theory of mind, a particular type of mental state (e.g. pain) corresponds with a particular type of physical state (e.g. C-fiber firing). This theory can be contrasted with, *a functionalist theory of mind* (the dominant theory), which says that a particular type of mental state corresponds with a particular type of function (e.g. yelling ouch and withdrawing). If the identity theory of mind would turn out to be correct (which is not likely - see Schneider, n.d.), and mental states are the mark of understanding, then we could be able to detect the appropriate mental states through their *corresponding physical states*. Unfortunately, we would still be at a loss at discerning whether someone understands anything at all until we figure out *which physical states* correspond with the appropriate mental states. Furthermore, we would still need to figure out *which mental states* are valuable, which brings us to the second problem.

The second problem is that there is no clear way to decide which occurrences are appropriate by referring only to those occurrences themselves. What is the correct mental state to understand the irrationality of the square root of 2? This is not only hard to characterise, but the avenues in which we

look to justify our choices are invariably motivated by the external affordances granted, not by the intrinsic value of the occurrence itself. What a group of subjects with understanding share (or overlap in) is the appropriateness of their abilities, not the appropriateness of their presumed states. Calling the occurrence in one's mind's eye "a proof" would be vacuous if what is seen cannot be put to any public use (or worse yet, incorrect use). Conversely, if we found out that someone exhibiting extreme competence had mental states different from what we have hitherto characterised as appropriate, we would broaden the scope of appropriate mental states, not deny her with understanding.

This brings us back to the possibility of being wrong about one's own mind. All of us can have difficulty expressing what kind of mental representations we are operating with. And even when we think we can, our presumptions about the processes of our minds are not always an accurate description about the actual processes that underlie them. Studies in mental rotation, for instance, have shown that while we think we can rotate and compare 3D Tetris-figures in our mind's eye, the competences we have do not line up with computer-systems that actually do what we think we do. (Dennett, 1993, c10) Confronted with such a situation, we are forced to rethink what we know about our own mental representations. This just goes to show that the kind of mental representations we infer is better marked by considering the competences they imbue than vice versa. The nature of the internal model is judged by external displays, not the other way around.

This does not imply that we must dismiss all claims made from a phenomenologically "internal" narrative. Our focus is on how people act, but how people act includes their speech acts, which includes speech acts about their own phenomenology. We must take these speech acts seriously (i.e. as genuine evidence of that person's abilities or lack thereof), but we're under no obligation to take their internal seeing-narrative metaphor literally (i.e. as correct claims about the mechanisms of the mind behind those abilities, the mental states).[16] Furthermore, it entails that characterising our understanding by putting a premium on our mental representations or operations puts in a position where we are vulnerable to misleading or mistaken ontological or epistemological claims about our minds and the way they work.

One way out of this problem would be to say that the appropriate occurrences are exactly those of which the public effects are certain appropriate abilities, a route taken by Wilkenfeld (2013a), for instance:

[16] For an exposition on the third person phenomenology (heterophenomenology) indicated at here, see (Dennett, 1993).

> "[O]ne's understanding x consists in large part of representing x in the right sort of way (...) I will contend that "the right sort of way" is best cashed out as a mental representation the possession of which enables certain abilities" (Wilkenfeld, 2013a, p. 1002)

But then our understanding-attributions are actually decided by the abilities and not the occurrences which supposedly lie behind them. In short, it cannot get around the problem that "the correctness of the internal model [or representation] is judged by external displays of understanding, not the other way around." (Ylikoski, 2009, p. 103). We may as well put the horse in front of the carriage, where it can do its work unencumbered.[17]

Having dispelled the mental state approach does not mean, however, that we cannot use the concept of mental states or representations as *instrumental postulates* (i.e. as a proposed explanation that binds the observable input and output), but then they are still "only hypotheses, models designed to explain, to sum up, what you observe" (Wittgenstein, 1953, p. 62). The dispelling is only meant to target mental representations as the *mark* of understanding, where understanding stands or falls based on what happens inside the mental sphere, such as what is in someone's mind's eye or other mind organ, conscious or not. Within each there is still a lot of freedom in how precisely we conceptualise the mark of understanding, but we have now seen the pitfalls of an approach that puts its premium on mental states directly. Furthermore, there are traces of this conceptualisation in philosophy, and it contributes to conceptual problems that could be avoided, and it leads to scepticisms about the ability account that are unsupported or inconsistent (as we shall see in Chapter 3, where I address objections against the ability account).

Synonyms

The literature makes frequent references to understanding involving "grasping" (e.g. Kvanvig, 2003; Trout, 2005; Khalifa, 2013, Grimm, 2011, 2014), or "seeing" (Zagzebski, 2001; Riggs, 2003), or "having" something, but what exactly is being grasped, seen or had (be it a representation, a proposition, an explanation, a relationship, a belief, a body of information,...) as well as what that entails (be it a sense, an attitude, an ability, mental access to,...) is something that varies depending on who uses the term, which means it still requires spelling out before it characterises anything, beyond supplying a (near-)

[17] Even if we put a premium on the abilities, is it not always some state that enables that ability? While it is true that abilities always have an implementation that allow them to manifest, we don't need to refer to that implementation, and when we do, we can do so without letting our ontology of that process do the demarcating. We talk of a car being driveable because it has the disposition to drive, not because it has an appropriate state that enables driving. We can talk about the implementation of that driving, such as various types of engines (or engine states) or mechanical procedures (or states), but what is shared among all these driveable cars is not a state, but that they drive.

synonym or metaphor. I'll now consider some of these synonyms or metaphors and the obscurification or pitfalls they can lead to.

In a literal sense, the concept of "grasping" is about manually seizing (and possibly manipulating) something, and the concept of "seeing" is about a visual process that stretches from the eye to the brain. It might be readily conceded that "seeing" here is not primarily about the literal visual process performed between eye and brain, or even primarily about a corresponding mental process performed by the mind's eye (and the mind's brain?[18]). And it might be readily conceded that "grasping" is also not primarily about a literal seizing performed by hands, or primarily about corresponding mental gestures with the mind's hands. So their usage could be metaphorical - but that leaves us with the question of how metaphorical the usage is, and what the value of the metaphor is. Quite often, the work that the metaphor is supposed to be doing (and, equally important, what it is *not* supposed to be doing) is not made explicit - which is unfortunate given that it is intended to lead us to a more explicit concept of "understanding" (Gordon, n.d.). Now, I don't mean to discredit entire accounts of understanding based on the use of (near-)synonyms or metaphors, but I will show how the use of synonyms or metaphors in the endeavour to mark understanding can obscure the concept and/or (mis)lead us to the same pitfalls discussed earlier.

Sometimes, it seems the concept of "seeing" lacks explicit elucidation. Zagzebski (2001) says understanding involves "*seeing* how the parts of that body of knowledge fit together, where the fitting together is not itself propositional in form" (p. 244, italics added), and that it involves "mental representations" (p. 241) which she thinks "will likely include such things as maps, graphs, diagrams, and three-dimensional models in addition to, or even in place of, the acceptance of a series of propositions." (idem), but it is unclear how literal she takes "seeing" and how phenomenological she takes mental representations to be.[19] Similarly, Kvanvig (2003) uses the term "grasping" directs the grasping to coherence relations, and qualifies it as "internal" (p. 192), but he never spells out what grasping involves, so it is hard to know what it is literally doing or metaphorically pointing to. (Gordon, n.d.)

> "Understanding requires the grasping of explanatory and other coherence-making relationships in a large and comprehensive body of information. One can know many

[18] A problem which Dennett addresses as the "myth of double transduction". (See Dennett, 1998a)

[19] She does mention that understanding "is a state that is constituted by a type of conscious transparency" (p. 246), meaning that, unlike knowledge, you can't understand without understanding that you understand, which is a denial of tacit understanding, but not necessarily an endorsement of mental "seeing".

> unrelated pieces of information, but understanding is achieved only when informational items are pieced together by the subject in question." (Kvanvig 2003, p. 192)

Grimm (2011, 2014) starts in a similar place, arguing that understanding (which is the same as non-propositional knowledge of causes) consists of "seeing" or "grasping" a modal relationship (as opposed to a proposition). The verb "seeing" is explicitly acknowledged as a metaphor, but what work is it doing here? Grimm, at least, is more explicit about that:

> "What the metaphor of "seeing" seems to involve, then, is something like an apprehension of how things stand in modal space (...) Just as, in seeing with one's eyes, one takes in or apprehends how things stand in the physical terrain, so too the basic idea here seems to be that in "seeing" with the eye of the mind, one takes in or apprehends how things stand in the modal terrain: one apprehends what cannot be otherwise, or how certain changes will lead, or fail to lead, to other changes." (Grimm, 2014, p. 334)

The "seeing" seems to be a bodily metaphor for a correlate in the mind's eye. But if one can passively witness changes in physical terrain, then can't one also witness how certain changes lead to other changes just as passively? What makes subject A understand better than subject B, is not that subject A has a mental sensation of a relationship that subject B doesn't have, but that (s)he can exploit such a relation.[20] Furthermore, if "seeing a relation" is read as synonymous to "exploiting that relation", we are already moving away from putting representations center-stage in favour of abilities. The passivity is addressed by Grimm elsewhere, where he points out the manipulist nature of this "grasping".

> "[in the manipulist] sense, mentally to grasp (...) a structure would therefore seem to bring into play something like a modal sense or ability—that is, an ability not just to register how things are, but also an ability to anticipate how certain elements of the system would behave, were other elements different in one way or another." (Grimm, 2011, p. 89)

We now have "modal ability", consisting not merely of "apprehending" or "registering", but also "anticipating". What is this modal ability? According to Grimm (2014):

[20] This is also the reason that de Regt (2009) invokes a skill condition. More on that later.

> "On our proposal, "seeing" or "grasping" would count as a kind of ability, because the person who sees or grasps [modal relations] will characteristically have the ability to answer a variety of what James Woodward (2003) has called "What if things were different?" questions." (Grimm, 2014, p. 339)

Is the modal ability the ability to answer what -if questions (it doesn't seem that way, given that the subject will only characteristically have that ability) or is this ability a symptom of another modal ability that occurs inside the mental realm? In trying to demarcate understanding, it is unclear what work the occurrence inside the mental realm is doing and what work is done by the ability to answer what-if questions. It is not entirely clear whether we've made the concept of understanding more explicit or the metaphor more complex.

Hills (2009, 2015) uses the metaphor of "grasping", calling the metaphor "an extremely important one" (2015, p. 4), but also "not very clear" (idem). She explains the meaning in a similar way to Grimm:

> "When you grasp a relationship between two propositions, you have that relationship under your control. You can manipulate it. You have a set of abilities or [intellectual] know-how relevant to it [incl. intellectual know-how], which you can exercise if you choose." (Hills, 2015, p. 4)

But what that intellectual know-how entails, she spells out as concrete abilities (which we'll see in Section 1.3), although she doesn't say whether they mark understanding or are merely a symptom of it. She explicitly leaves it open: "Is understanding why (partly) constituted by these abilities? Or is it the ground of these abilities? I favour the former account, but it is one of the questions that I will leave open here." (Hills, 2015, p. 5)

Khalifa (2013) argues that grasping an explanation is central to understanding and necessarily entails true beliefs (of the form q explains p) must be the result of exercising reliable cognitive abilities, which will involve evaluating (or discriminating between) explanations. This needs further explication by describing what it means to have a belief and what exactly is involved in evaluating an explanation, which is related to the question of what makes it a *cognitive* ability.

It is worth making a quick side-note here about the word "cognitive ability," which used by other authors as well. What sort of modifier is "cognitive" in "cognitive ability"? Does it specify abilities as

(i) an ability that can be categorised as primarily cognitive (as opposed to, say, tennis, which can be categorised as a primarily physical ability), as (ii) produced by a cognitive entity (as opposed to provided by the environment)[21], or as (iii) private performances taking place within the secret realm of the mental (as opposed to potentially[22] taking place as public acts)? Wilkenfeld (2013a) could be attributed with the third, for he says "Understanding is a cognitive achievement (...) [which] must instead consist of an ability to manipulate some mental correlate of the understood object" (p. 1003). Greco (2007), on the other hand, uses the term "intellectual ability" in a similar way, but clarifies the nature of "ability" as "dispositional properties that display a characteristic structure" (p. 68) which clearly places abilities outside of the mind. Pritchard (2014) is not as clear on this point. He explains that an archer has an achievement (e.g. hitting the bullseye) if she exercised her ability (e.g. archery ability) and it was a success (e.g. hitting it) mainly due to that exercise as opposed to due to other factors (e.g. wind). So it also goes with the cognitive: someone has a cognitive achievement (understanding) if she exercises her cognitive ability (reliably forming a true belief) and it was a cognitive success (a true belief was formed) due to that exercise (and not, for instance, by trusting the testimony of an expert). Pritchard's use of the term evokes the first reading of cognitive ability (it distinguishes the *cognitive* ability to form true beliefs from the *physical* ability to hit a bullseye) and the second (it serves to distinguish the credit of the success due to the agent's ability or something external), but it is unclear whether he supports the third - because the ability of having or forming a true "belief" is itself in need of clarification. Pritchard's account shifts the issue to the nature of belief.

It seems uncontroversial to suggest that understanding involves beliefs. Most philosophers make reference to the notion of belief in some way (see e.g. Grimm, 2011; Khalifa, 2013; Kvanvig 2003; Hills 2009; Pritchard, 2014). But belief is not quite such an uncontroversial term. It is far from obvious that beliefs exist (see section 1.4), and if they do, it is far from obvious what interpretation is the most useful one (see Dennett, 1990, c5 for a rundown on the interpretation of the term and their problems). Once again we are faced with a term that, in the context of characterising understanding, moves the target rather than marks it. Is a belief a mental state, as is suggested by the traditional notion of propositional attitudes (which is still very open for interpretations about what sort of entity a proposition is, where we find it, and how we take an attitude towards it[23]) or a mental representation (with the problems that come with it) (Schwitzgebel, 2019), or is the notion of belief a claim about performance, as is suggested by the dispositional or interpretationist reading? In the case of the latter,

[21] The issue of subject-demarcation is one we'll come back to in Part II, most notably in Chapter 4.

[22] Just to be clear: no one is denying that people can keep their abilities to themselves, but this is a radically different thing from claiming that their abilities are inherently *located* in the realm of the private.

[23] Some of them (mentioned in Dennett, 1990, c5) include "grasping" a proposition - stop me before I get dizzy.

an ability approach (which also rests on performance) might have saved us a lot of conceptual legwork. We'll return to the notion of belief briefly in section 1.4 and more extensively in Chapter 4. For now, we'll put the term to the side in favour of a more steady target.

1.3 Defending an Ability Account

An alternative proposal to mark understanding is to place a premium on the presence of abilities (e.g. Avigad, 2008 for mathematical understanding; de Regt & Dieks, 2005; Hills, 2009 and Ylikoski, 2009 for scientific understanding). This is the approach I will be defending as the most sensible and fruitful. I will start by roughly characterising my approach to understanding thusly:

> 'S understands X' corresponds to 'S possesses sufficient abilities appropriate to X in context C.'

While it appears to be a simple characterization, there is still a lot to unpack here. Every word in this characterisation requires further elaboration: which "abilities" are appropriate, how many are "sufficient", how one can "possess" them, which types of entities "S" can do so, as well as what abilities are "appropriate", how to discern them and which role the "context" of the attributor and the circumstances of the subject play in this. Each of these questions will be considered in the course of this dissertation, but for the purpose of defending the ability-approach as a mark of understanding, I will give a few clarifications in this section as they pertain to the mark of understanding. What is of prime importance here is that if we characterise understanding through abilities, then the successes of an understanding attribution will need to come from the subject's appropriate acts or performances.[24] Conceptualising such an ability account will take us beyond the mental (which will force us to focus on performances, with the virtues that that entails), beyond single performances (which will force us to consider the concept of ability, which I will conceptualise as multi-track behavioural profiles), and even beyond a single ability (which will force us to consider the notion of the *appropriate* set of abilities, which I will conceptualise as an appropriate behavioural profile).

Benefits of an Ability Account

Under the ability account, attributions of understanding stand or fall with how a subject acts (i.e. performs). As will become clear, this does not involve reducing everything to stimulus-response

[24] I will use both "act" and "performance" as synonyms for appropriate behavior: act as in "performing an act", not as in "voluntary behaviour"; performance as in "a successful act", not as in "presenting a fiction". Which one I use will depend solely on which seems more appropriate to avoid the misleading connotation.

descriptions, but it does involve that any appropriate attribution of understanding bottoms out in claims about potential outward performance, and not anything that lies behind those performances.[25] There are a number of benefits or virtues to this approach, and I will go over some of them in this subsection.

Firstly, we are side-lining the role of feelings we associate with understanding without necessarily discarding them as epistemologically irrelevant. When we are specifying what marks understanding, we can refer to appropriate abilities without reference to the emotions or sensations that accompany, motivate or implement them. At the same time, this does not discard the role of feelings outside of the purview of epistemology. To the extent that feelings do play a role in the implementation of a subject with abilities, goal in the subject's pursuit of them, or guide in the process of attaining them, they are also (with varying degrees) relevant to epistemology. An ability approach acknowledges this.

Secondly, by putting a premium on abilities, we are avoiding some of the problems that plagued the accounts we discussed earlier. If a subject's mental states are inferred by the acts of a subject (including speech acts about the subject's own phenomenology), then those mental states can only be discerned indirectly, if at all. But a subject's abilities are literally comprised of such acts, which means that the acts they display serve as direct (though incomplete[26]) evidence of a subject's abilities. Next, it is also easier to pinpoint the appropriate abilities because we can refer directly to the appropriate acts themselves, as opposed to looking for the acts which correlate with the appropriate mental state. Additionally, determining which abilities are appropriate is more to the point because abilities can have intrinsic values (e.g. we value prediction for its own reward). Furthermore, justifications of understanding tend to boil down to abilities when asked to justify them (even if they are initially phrased in the mental state narrative). For instance:

> "Suppose I tell you that my friend Paolo understands group theory, and you ask me to explain what I mean. In response, I may note that Paolo can state the definition of a group and provide some examples; that he can recognize the additive group structure of the integers, and characterize all the subgroups; that he knows Lagrange's theorem, and can use it to show that the order of any element of a finite group divides the order of the group; that he knows what a normal subgroup is, and can form a quotient group and work

[25] Chapter 3 is entirely dedicated to showcasing how a variety of cases bottom out in abilities as conceived through my account.

[26] That a single act is insufficient for constituting an ability is something that could underlie an objection against the ability-account. This problem is discussed in Section 2.3 as well as further addressed in the examples of Chapter 3.

> with it appropriately; that he can list all the finite groups of order less than 12, up to isomorphism; that he can solve all the exercises in an elementary textbook; and so on. What is salient in this example is that I am clarifying my initial ascription of understanding by specifying some of the abilities that I take such an understanding to encompass. On reflection, we see that this example is typical: when we talk informally about understanding, we are invariably talking about the ability, or a capacity, to do something." (Avigad, 2008, p. 321)

Another benefit of the ability approach is that the notion of tacit understanding is given more room to flourish. People have been shown to respond appropriately to situations without being able to articulate (or even register) what they are doing. For instance: people are able to detect subtle variations in human appearance and behaviour without being able to tell what they detected, people are able to extract temporal patterns that underlie a train of events without even having noticed they did so, people are able to learn without understanding how they learn, or even knowing that they have learned (i.e. implicit learning). (Reber, 1989) And people are able to form judgements about the quality of an explanation without knowing how they do so. Here "the contrast between our ability to make [these judgements] and our inability to describe them on the basis on which we make them is particularly stark." (Lipton, 2009, p. 60). Such epistemic competences have been called *tacit.* Whether they deserve the label of understanding (or knowledge) is where internalists and externalists (be it about knowledge or understanding) differ. Externalists believe reliable competence is enough, but internalists claim that one cannot understand unless one can also articulate, justify or explain one's understanding. Zagzebski (2001) explicitly makes the latter claim:

> "Understanding, (...) is a state that is constituted by a type of conscious transparency. It may be possible to know without knowing that one knows, but it is impossible to understand without understanding that one understands. (...) understanding is a state in which I am directly aware of the object of my understanding, and conscious transparency is a criterion for understanding." (Zagzebski, 2001, p. 246)

Under a state account, this seems intuitive, because how would we know there was a "mental state" if its presence or constitution wasn't in some way "accessible" to the subject.[27] Nonetheless, conscious transparency seems too strong a requirement for understanding. Mathematical competences, for instance, are valuable whether they are consciously deliberated or subconsciously brooded. Pritchard

[27] Unless it is inferred from the subject's behaviour, in which case the premium once again shifts to abilities.

(2009) also thinks Zagzebski takes the transparency condition too far, but agrees with her that when one has understanding, as opposed to knowledge, "it should not be opaque to one that one has this understanding – in particular, one should have good reflectively accessible grounds in support of the relevant beliefs that undergird that understanding." (p. 39) Nevertheless, it is not clear exactly why this should be so, nor that it should be so. The suggestion seems to be that understanding, unlike knowledge, requires more competence, and the internalist condition is supposed to ensure this. But the internalist condition thus also discredits all other forms of competences, which is unfortunate. Grimm (2016) argues against "articulacy" as a necessary condition for understanding by appealing to understanding in children and animals. We can make similar points about experts. While mathematicians can prove and recognise a proof when they see one, it is notoriously difficult to say what it is that experts do while proving or what it is that they recognise when they see a proof. Consider this quote by the mathematician Reuben Hersh[28]:

> "When you're a student, professors and books claim to prove things. But they don't know what's meant by 'prove'. You have to catch on. Watch what the professor does, then do the same thing. Then you become a professor, and pass on the same 'know-how' without 'know what' that your professor taught you" (Hersh, 1997, p. 50)

While this may be cause to say that understanding of proof is incomplete, we'd have a strange account of understanding if it entails that mathematicians don't understand the concept of proof. Under an ability account, it becomes clearer that the absence of the ability to articulate or explain one's understanding does not have to discredit other abilities; rather it marks a lower degree of understanding (more on degrees in Chapter 2). Furthermore, what is missing is simply more abilities, namely the ability to articulate what you understand (and other related abilities that may rely on that ability). Furthermore, the distinction can be made clear with the terms tacit and theoretical understanding (see Ylikoski, 2009) or implicit and explicit understanding (see Hills, 2015).[29]

The ability account also sidesteps the problem of excessive or infinite encoding. Allow me to explain. If one subscribes to the idea that there is a grasping relation (e.g. an attitude) towards an object of understanding (e.g. propositions), then a complete understanding will, for instance, involve an attitude towards *all the relevant propositions*, "the ideal understanding text" (Van Camp, 2014, p. 108,

[28] See also Thurston (1998) and Davis & Hersh (1998).

[29] Furthermore, an internalist philosophy may already have made up its mind about what is "internal" and what is not, disallowing, for instance, external representations from playing a role in epistemology - regardless of how large a role they play in science. (Kuorikoski & Ylikoski, 2015) More on the internal/external divide in Chapter 4.

based on Railton, 1981), which "consists of a complete framework of all possible propositions about the phenomena, including their various relationships" (p. 108) This "ideal understanding text" may grow exponentially large with each relation, or infinitely large due to Carrollesque recursion[30], placing some strain on the physical or mental space that needs to encode it all. The ability account, by contrast, doesn't require explicit encoding of all the relevant propositions that can be constructed, as long as one can keep responding appropriately.

There are further issues with encoding states that the ability account avoids. Think of the following: How specific or compartmentalised does a state need to be for understanding to be attributable? Once someone understands, is there a mental or physical state which is permanently present as long as the person can be said to understand? Does that person lose the understanding as soon as she's no longer in the previous state? Is there a compartment of that person's whole mental or physical state that covers the particular understanding (and how many state-slots can there be before someone's whole state is "full"?) Or is the state transient and do you stop understanding as soon as the mental or physical state has changed? Or do we have to refer to potential states, (making it even more difficult to ascertain whether someone understands or not). What it means to possess a mental representation or have access to the state is just another extra step of vagueness we can do without. What subjects with understanding share (or overlap in) is their abilities, not their states.

A last virtue of the ability account that I will mention is its improved resistance to chauvinism. Within sense or state accounts, one could side-line all entities one isn't keen to attribute understanding to (such as other ethnicities, genders, or species) by marking out an inevitable difference in physical state or constitution[31], and simply discrediting a subject's postulated mental state (or simply denying the subject with the "mental" modifier altogether[32]) without specifying what makes the difference a relevant one. This form of chauvinism does not have to be explicit or intentional to have an effect. But even an implicit chauvinism would be much harder to substantiate if one has to mark a valuable difference in scientific performance than in physical or mental constitution. Of course, the ability account is not fully chauvinism-proof. It would still be possible to deny the value of certain acts or performances for chauvinistic reasons, but to do so one will be faced with the more demanding task

[30] See Lewis Carroll's (1895) *What the Tortoise Said to Achilles.*

[31] For instance, one could use it to only allow certain physical constitutions or a particular physical make-up (e.g. "to understand, one needs x percent water, y percent carbon, z percent salt"). Think of the claim that computers can't understand because they're not organic, or Searle's (1980/1985) response to the water pipe simulation argument.

[32] E.g., "humans can grasp meaning, computers can only pretend to" or "humans are conscious, but an artificial replication would be a zombie."

of convincing a scientific community which acts to (not) value than which member's mental or physical constitution.

But in spite of its many virtues, there are still some causes for scepticism and these do need to be addressed. Firstly, one may think that the ability account has no way to talk about competence without performance (i.e. dormant competences) or one may object that a single performance is no guarantee for true competence. Both of these worries (and many more) can be addressed (which I will do in Chapter 3) by further developing the notion of "ability", so let's.

Beyond an Act & Ability

The ability account, I've already argued, demarcates understanding based on how a subject acts (i.e. performs), and whether those acts can be seen as appropriately successful. This is true for all abilities, not just epistemic ones. If Hamlet claims he is able to tell a hawk from a handsaw, then the validity of that claim rests on whether Hamlet correctly distinguishes one from the other. Likewise, whether a meteorologist is able to predict the weather stands or falls with how she fares in the predictions she makes. One reason why it may still seem appealing to refer to private occurrences beyond a person's performance is because it seems plausible for the same act to be performed with and without understanding – for instance, by sheer luck, rote memorisation, or blind rule-following. This seems to imply that the difference-maker for understanding lies not in the performance, but in something beyond it. But we may readily concede this without thereby having to withdraw into a secret world.

Firstly, "ability" is a modal predicate, concerned not just with how things are or have been, but how things could be. Breakable glass may never shatter, a meteorologist may never predict the weather and Hamlet may get through the entire play without ever distinguishing a hawk from a handsaw. And yet this alone is not sufficient to deny any of them with being able to shatter[33], predict the weather or distinguish a hawk from a handsaw. We can address this by opening up the range of circumstances under which one would perform appropriately.

> "[A]bilities in general are functions of success in relevantly close possible worlds. In other words, to say that someone has an ability to achieve X is to say that she would be successful in achieving X in a range of situations relatively similar to those in which she typically finds herself." (Greco, 2000, p. 13)

[33] One may wish to object that glass does not have the "ability" to shatter, but is merely disposed to. We'll address this concern later in this section.

So it seems abilities involve successful performances under a range of circumstances, regardless of whether these circumstances actually obtain. (We'll discuss the modality of abilities in more detail in Section 1.4). This means that when we ascribe someone as having an ability, we are using evidence of past and present performance to make an estimation of the quality and range of those performances under the salient circumstances. Making a lucky guess may be something one gets away with under the precise circumstances under which you got lucky, but luck quickly runs out if you are tested under different circumstances. So, the problem with sheer luck is not the success under a singular circumstance, but the range of failings under others.

Furthermore, the use of the plural in "abilities" is not incidental. Understanding involves more than just having a single ability appropriate to the object of understanding. I can easily memorise a correct response to a certain question (or even a few of them) without understanding what it is that I'm saying. The problem with such a single-track ability is not that it should be discredited, but that the scope of abilities is too narrow, way too narrow. Answering questions according to a set script may be something one can get away with in tests that happen to only involve the memorised questions (giving a misleadingly good impression), but this will quickly fail once tested for other (or related) appropriate abilities[34] (e.g. correct courses after a setback, explain in different words, apply it in a practical circumstance, make an analogy, instruct others, criticise incorrect practice, predict the outcome of observed lapses, answer what-if-things-had-been different questions,...). It is not the narrow *success* of rote memorisation, but the wide *failing* it entails that makes for a poor understanding.

The same is true of blind rule-following, what Skemp (1976) calls "rules without reason" (p. 20). It is not accidental that the adjective "blind" seems fitting. It is because any deviation or extension beyond where the rule leads, would leave the subject in the dark. It is not the narrow success of rule-following that is the problem, but the wide failing it entails. If we draw the scope of understanding (which we'll flesh out in Section 2.1) to include all the relevant abilities, then it becomes increasingly difficult to motivate why a subject displaying the wide range of relevant abilities would *not* merit an understanding-attribution.

[34] Where one ability begins and another ends will largely depend on how wide a net one is attempting to cast. The ability "to multiply two single digit numbers" clearly casts a wider net than the ability "to multiply 5 with 6". Some abilities don't have an obvious net-size. Consider the ability "to prove that the square root of 2 is irrational". Is this ability composed of the act, in the salient circumstances, of giving a proof, along with other acts (e.g. criticizing incorrect steps, predicting outcomes of observed lapses,...) or do these other acts constitute different abilities? Providing an answer is not only difficult, but will largely depend on how you phrase the ability in question. I will not concern myself here with providing a universal demarcation criteria for what constitutes a single ability. As long as one is with me in the claim that understanding casts a wide net, it doesn't matter to me whether it's because understanding captures many multiple abilities, or because abilities capture many acts. I'm more inclined to go for the former, for reasons that will become clear in Part II, but I'm happy to leave the precise size of the nets open or fuzzy.

There is an interesting example of how a mental state narrative can mistake conceptualisation for demarcation. Baumberger et al (2016), while giving an overview of different interpretations of grasping, presents us with the following situation:

> "Suppose that a climate scientist explains to her young son that the global mean surface temperature has massively increased since the middle of the 20th century because of increasing greenhouse gas concentrations. Since she is right and her son has good reasons to believe her explanation, he may be said to know why the global mean temperature has increased. But he does not seem to understand why. When asked why this is so, all he can do is to repeat his mother's explanation. The problem seems to be that he does not really grasp the explanation." (p. 12)

Does the diagnosis of "he does not really grasp the explanation" mark the real problem any more than what was previously stated? If we can note that all the boy can do is repeat his mother, did we not already pinpoint the problem? The job of grasping then, is in conceptualising the problem, not in marking it.

So far we've only been talking about performance in the most neutral sense, as a happening or behaviour. When we think of abilities, however, we're not thinking of glass having the ability to break in the same way that a meteorologist has the ability to predict the weather. In the latter we're thinking of a person having the skill to do something (i.e. the successful performance) and the option to engage that skill (i.e. the power or freedom to act). She will employ the skill only when she chooses to (Moore, 1912) or tries to (Fara, 2008). This may provide an appealing reason to retreat *behind* the act and look for the right kind of state that makes it a power over a mere reflex. But the difference between a reflex and a power lies not *behind* the act, but beyond it. The common intuition, for instance, is that glass is disposed to break, but doesn't thereby have the power to break. While I can appreciate the distinction in this case, must we therefore also say that the ability to calculate is different for humans than it is for calculators? Furthermore, does that distinction really matter in itself, or is the distinction related to the valuable differences in performance between a mathematician and a calculator? When I'm interviewing someone for a job based on her abilities, I'm not interested in whether she is "able to X" in the sense that her doing X was preceded by some appropriate mental state (or worse, some libertarian kind of non-determined free will[35]), I'm interested in whether she will do X with the desired

[35] For a defense of free will that is compatible with determinism and does not rely on an inaccessible mental realm, see (Delarivière, 2016).

level of sensitivity and when it is appropriate. The same holds for doing something "conscientiously". We are not peering *behind* the act to see whether it was performed with the appropriate mental state, but looking *beyond* the act, to see whether it was performed with the required level of sensitivity and quality. Acting conscientiously is not doing two kinds of things: being conscientious and acting, it is doing one kind of thing (i.e. acting) *well*. (Ryle, 1949/2000) For this reason, I won't distinguish between acts in the dispositional sense and acts in the power sense, because I believe they are distinct in degree, not kind.

In short, when we attribute people with understanding, we are not making untestable inferences to any secret phenomena which are forever out of our reach and judgement, but we are gauging the appropriate multi-track behavioural profiles of an understanding subject. So it is true that our understanding-attributions go beyond single acts, but this is not going beyond as in going *behind* them (to occurrences which are impractical or impossible to discern and don't themselves contribute anything of value), but beyond as in considering what people could and would do, namely their abilities - which we can discern, even if never fully.[36] (Ryle, 1949/2000) We may, instrumentally, speak of mental states, but we do so on the basis of abilities, and not the other way around.[37] Hence, it is the abilities that actually mark the understanding.

Brand of Abilities

Of course not just any ability is relevant to any understanding. The ability to make a good cup of tea is not indicative of understanding the general theory of relativity. Which brand of abilities are "appropriate" to understanding has a variety of candidates, depending on who you ask, what kind of understanding they are meant to capture, and what the object or field of understanding is. In this subsection, I will be discussing several of the candidate brands or kinds of abilities offered up in the literature as appropriate for understanding. Then I will offer up my own conceptualisation of the (contextually) appropriate dimensions and kinds of abilities in Chapter 2. So which kinds of abilities have been offered up as appropriate? Let's start with considering a popular distinction between three general kinds of understanding attributions:

(1) Lindsay understands *that* there is a housing crisis

(2) Olly understands *why* there is a housing crisis

(3) Natalie understands the housing crisis

[36] See Chapter 2 for a closer look at the dimensions and degrees that make up understanding.

[37] See Chapter 4 for a closer look at how one interprets beliefs from acts.

These three kinds of attributions have often been distinguished in the literature, beginning with Kvanvig (2003), as (1) understanding-that or propositional understanding (2) understanding-why or atomistic understanding and (3) objectual understanding, respectively. Although the distinction is widespread in use, it is not without its issues. The first kind of attribution, propositional understanding, has been criticized (e.g. Pritchard, 2010, Gordon, 2012; Grimm, 2016) as being indistinguishable from either attributions of propositional knowledge (if it only involves knowing a single proposition), or atomistic and objectual understanding (if it involves more). Furthermore, Grimm (2016) believes the distinction between the second and the third understanding is overstated, and largely due to a difference in focus or scope, rather than kind.[38] I'm inclined to agree (which is why in Chapter 2 I will offer an alternative approach to distinguishing kinds of understanding). Nonetheless, it is worth bringing up to know which focus or scope an author has in mind when they're offering up the necessary abilities.

Even though it is unclear whether they should count as symptom or trait, Grimm's (2011, 2014) suggestion for the crucial ability involved in atomistic understanding (and thus also for objectual understanding) is that of being able to answer what-if-things-had-been-different questions (based on Woodward, 2003). This suggestion is a powerful one that enjoys a lot of support. Hills (2009, 2015), focusing on understanding-why in particular, expands on Grimm. Although she left open whether abilities mark understanding or are mere symptoms of it, she does emphasise the ability to "to treat q as the reason why p, not merely believe or know that q is the reason why p" and spells out what that entails with a concrete list of abilities that covers not just answering what-if-things-had-been-different questions, but also a series of explanatory abilities, namely:

> "(i) follow some explanation of why p given by someone else
> (ii) explain why p in your own words
> (iii) draw the conclusion that p (or that probably p) from the information that q
> (iv) draw the conclusion that p' (or that probably p') from the information that q'
> (where p' and q' are similar to but not identical to p and q)
> (v) given the information that p, give the right explanation, q;
> (vi) given the information that p', give the right explanation, q'" (Hills, 2015, p. 4-5)

[38] "For both Kvanvig and Pritchard, "objectual" or "holistic" understanding has to do with our grasp of large chunks of information, especially as they relate to topics or subject matters. Understanding-why or atomistic understanding, by contrast, is focused on some particular state of affairs: understanding why the cup spilled, for example, or why Fred did poorly on his exam." (Grimm, 2016, p. 254)

Kuorikoski & Ylikoski, (2015; Ylikoski, 2014), also take inspiration from Woodward's what-if-things-had-been-different questions, but they put the abilities front and centre, calling them counterfactual inferences. They include instances of predictions, control and explanations under what-if circumstances. Wilkenfeld (2013b) pushes back on this characterisation by pointing out that counterfactual inferences do not help us with *necessary* truths, as is the case in mathematics, where "all (or at least almost all) of what one can understand involves necessary truths and the relations they bear to each other, and so there is no counterfactual dependence involved at all." (p. 101) This may be slightly too strong a pushback. There are multiple answers to "what if"-questions that don't rely on one to bend necessary truths. For example, in the case of the irrationality of the square root of 2, one can ask: what about the square root of 3? See (Frans & Weber, 2014) for more in depth examples of what-if-things-had-been-different questions and answers in mathematics. But the point may be taken that answering what-if-things-had-been-different questions or counterfactual inferences may not satisfy all needs.

Speaking of mathematical understanding: Avigad (2008), focusing on mathematical understanding, takes what he calls a functionalist approach to understanding. Taking lessons from Wittgenstein, this involves characterising understanding through the relevant abilities. For proof, he offers the following list of (kinds of) abilities involved:

> "• the ability to respond to challenges as to the correctness of the proof, and fill in details and justify inferences at a skeptic's request;
> • the ability to give a high-level outline, or overview of the proof;
> • the ability to cast the proof in different terms, say, eliminating or adding abstract terminology;
> • the ability to indicate 'key' or novel points in the argument, and separate them from the steps that are 'straightforward';
> • the ability to 'motivate' the proof, that is, to explain why certain steps are natural, or to be expected;
> • the ability to give natural examples of the various phenomena described in the proof;
> • the ability to indicate where in the proof certain of the theorem's hypotheses are needed, and, perhaps, to provide counterexamples that show what goes wrong when various hypotheses are omitted;
> • the ability to view the proof in terms of a parallel development, for example, as a generalization or adaptation of a well-known proof of a simpler theorem;

> • the ability to offer generalizations, or to suggest an interesting weakening of the conclusion that can be obtained with a corresponding weakening of the hypotheses;
> • the ability to calculate a particular quantity, or to provide an explicit description of an object, whose existence is guaranteed by the theorem;
> • the ability to provide a diagram representing some of the data in the proof, or to relate the proof to a particular diagram;
> And so on" (Avigad, 2008, p. 327-328)

But the most extensive ability account (and most extensive understanding account in the literature of the philosophy of science, full stop) is the contextual account of scientific understanding provided by de Regt (& Dieks, 2005; 2009). de Regt starts by criticising the inadequacies of beliefs or theories if they can't be put to some use. As such, he places a *skill condition* on understanding. His focus is on scientific understanding, so his criteria also involves explanations on the basis of a scientific theory. As criterion for understanding a phenomenon (CUP) scientifically, he offers:

> CUP: "A phenomenon P is understood scientifically if and only if there is an explanation of P that is based on an intelligible theory T and conforms to the basic epistemic values of empirical adequacy and internal consistency." (de Regt, 2009, p. 92)

Note that this criterion is not a pragmatic one, for it makes no reference to a subject that understands. The pragmatic element lies in another term, namely the "intelligibility" of a theory (CIT). For this, he offers a sufficiency criterion (at least for disciplines which formulate theories in mathematical terms, such as the physical sciences):

> CIT: "A scientific theory T (in one or more of its representations) is intelligible for scientists (in context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculations." (de Regt, 2009, p. 102)

In short, one could say that CIT involves what it means to understand a theory, whereas CUP involves what it means to say a theory is scientific and can be understood.[39] It may seem inviting to read that what understanding involves in his account is to "have an explanation," because CUP requires merely that an explanation exists and de Regt doesn't quite specify the link between the explanation and its

[39] Personally, I would read CUP not as a criterion for understanding, but as a criterion for scientific adequacy. CIT, on the other hand, seems to me the true criterion of understanding: namely that the intelligibility of a theory grants us an understanding of the phenomena - provided the theory in question is scientifically adequate (passes CUP).

use in understanding. However, he does also say that a phenomenon is understood (scientifically) if one understands a scientific theory, and that understanding a scientific theory involves skill and abilities. This steers us clear of the notion that understanding would involve owning a scientific explanatory text or possessing a scientifically accurate mental representation. Together, CUP and CIT suggest that phenomena can be understood by a scientist if there exists an explanation based on an empirically adequate and internally consistent theory of which the scientist is able to recognize qualitatively characteristic consequences of the theory without performing exact calculations. Wilkenfeld (2013) gives a more technical interpretation of how to combine de Regt's two criteria:

> "UD2: A phenomenon P is understood scientifically if there exists a scientist S in context C such that S (in C) can explain P with T and S in C can recognize qualitatively characteristic consequences of T without performing exact calculations and the explanation of P by T meets accepted logical and empirical criteria." (Wilkenfeld, 2013b, p. 98)

A virtue of this particular ability account is that it allows intelligibility standards to vary along with the history of science and theories of scientific explanation. It leaves open "empirical adequacy", which allows it to vary with the progress (or history) of scientific methodology and the variation of methodologies that come in different scientific disciplines. I too will leave the criteria of "good science" to the context of attribution (see Section 2.2), so my account can also be considered as contextual. While our accounts take a different approach, I believe they end up in the same place with regards to its contextual and scientific nature. For instance, under de Regt's account, the situation with astrology is that astrological theories can be understood (passing CIT), but are not scientifically adequate (so doesn't pass CUP). (see de Regt & Dieks, 2005) Under my account, the problem of astrology depends on the context of attribution[40] (more on this in Section 2.2).

de Regt's account is not without its criticism. His theory requirement in particular has been attacked as too strong because we don't always need a theory to understand (Kelp, 2015). de Regt's focus, however, is meant to be on scientific understanding, not everyday understanding of science. de Regt's focus on theories does constrain his account, but this constraint seems fair given that he's not focused on attributions of individual scientists, but the progress of the sciences as a theory-generating discipline (even while not betraying the pragmatic nature of understanding). Because I talk about understanding in a broader sense, I will cast a wider net. Newman (2012) objects along similar lines as

[40] For instance: either the context of attribution values predictions and astrological theories will come up short, making understanding attributions inappropriate if there is no predictive power (even by their own standards), or their context devalues predictions (or a certain kind of prediction), making their context of attribution unscientific.

Kelp, saying that even if we have a theory, we don't always have the highly theoretical skills to use one, and that if it is only about qualitative consequences, then it is not clear what makes it scientific understanding. (Newman, 2012) I think Newman is reading de Regt either unfairly narrow or unfairly broad, without allowing anything in between. That being said, I have my own criticism of de Regt related to his theory-focus. We'll come back to that in Section 3.3, but I can say, in short, that I have issues with "characteristic consequences" because he adds "without exact calculation" even though there is nothing particularly inappropriate about exact calculation. Presumably "without exact calculation" is added because one could calculate without knowing what one is doing, but the problem here is not in the exactness, but elsewhere (see the blind rule following objection in Section 3.3). Nevertheless, even if exact calculation is contestable in its appropriateness, recognising qualitatively characteristic consequences is not.

There are multiple candidates for multiple kinds of abilities in multiple fields, including: recognising qualitatively characteristic consequences (de Reg & Dieks, 2005), making counterfactual inferences in the contexts of manipulation, prediction and explanation (Ylikoski, 2009), give explanations (i.e. answering explanation-seeking questions) (Ylikoski, 2009), answering what-if-things-had-been-different questions (Woodward, 2003; Grimm, 2014), being able to evaluate explanations (Khalifa, 2013), relating knowledge to other knowledge (Van Camp, 2013), controlling the phenomenon (Yliksoki, 2009), specifying causal dependence (Ylikoski, 2014), reliably tracking dependency relations (Grimm, 2016), following an explanation given by someone else, explaining it in your own words, drawing the appropriate conclusion based on a given set-up, or vice versa (Hills, 2009), responding to challenges as to the correctness of a proof (or theory, or label), identifying key features, identifying the nature of the objects and questions, mustering the relevant background knowledge, exploring the space of possibilities fruitfully, and so on. (Avigad, 2008).

So which (type of) candidate is the correct one? My answer is that there is no single exhaustive candidate. I believe the scope of understanding to be quite wide (i.e. composed of, but not exhausted by any single set of candidates discussed in the literature) and contextually dependent on the many aims and values of the epistemic practice for whom the understanding-attributions is relevant. Each field has its own objects of understanding, each with their own needs for what they are supposed to satisfy (needs which may, furthermore, change over time). Each kind of understanding (be it for objectual, atomistic or propositional understanding, or otherwise) has its own focus or scope. While the commonality here is abilities, it is difficult to find a clear characterisation of a single type of ability that covers all the needs of the fields of epistemology and kinds of understanding previously

mentioned. So what I propose is that we keep constant what stays constant and then we can look for systematic ways to talk about what varies (which I'll do in Section 2.2, where we'll elaborate further on how to conceptualise the appropriateness of the abilities). Nevertheless, when it comes to the mark of understanding, there is but one sure commonality and that's the presence of the appropriate abilities.

1.4 Deriving Instrumental Concepts

Just because we put a premium on acts, doesn't entail that all our conceptualisations regarding understanding need to be framed in terms of acts in the same way that, for instance, behaviourism would have us do. Under behaviourism, all claims are essentially stimulus-response rules that do not allow us to postulate anything outside of observable behaviour. Behaviourism was superseded by functionalism, which still placed its premium on behaviour, but allowed instrumental postulates about internal states, but they were warranted only to the extent that they played a role in producing observable behaviour. Even in conceptualising understanding, there are quite a few concepts that aren't (directly) about observable behaviour or acts. I am not arguing to remove these from our conceptual toolbox all together (in fact, I will be employing a few of these myself), but I will argue that in each of these, the concept makes the most sense as indirectly targeting acts. Under my account, they are instrumental postulates, so they have explanatory or predictive value, and it is to the extent that they do, that we will employ them. I will be discussing three such concepts[41] that are seemingly not about observable acts - namely the modality of abilities, the meaning of objects and the mind of a subject[42] - but of which the warrant or justification still boils down to acts. These will help us conceptualise the quality of understanding (in Chapter 2) beyond mere lists of acts or abilities.

Modality of an Ability

First we have the *modality* of an ability. Modality is concerned with what could, must or cannot be the case. The modal concept that I want to focus on here is that of *counterfactuals*, which is about what "could or would have been" (Star, 2019). We have already established that for a subject S to possess an ability to act, it was not *sufficient* for S to act or have acted appropriately (because it may be luck), nor was it *necessary* for S to act or to have acted, because S having an ability to A doesn't mean S does A, but that S *could* do A. This "could," I'll argue, is best explained by a systematic way of dealing with the relevant "if" - the relevant circumstances where S does do A. Counterfactuals are a *conditional analysis* of the form "if A had occurred, then B would have" or "If A had not occurred, then B would

[41] I will be heavily inspired by the works of Daniel Dennett for all three of them.

[42] They are connected to the trait (T), object (X) and subject (S) of understanding respectively.

not have occurred" (Menzies, 2014). They have been developed in Lewis's (1973) counterfactual analysis with a focus on causal links, or Dennett & Taylor's (2002) possible world interpretation of counterfactuals to focus on the notion of "could have done otherwise." The focus here is not (primarily) on whether there is a causal link or whether a subject could have done otherwise, but on whether there exist circumstances where we would find our subject performing appropriately (and whether those circumstances are salient). From the perspective of counterfactuals, an ability can be conceptualised in the following way:

> *A counterfactual theory of ability:* S has ability to do A iff S successfully performs A in a set of salient (counter-)factual circumstances.

There's still quite a lot in this sentence to unpack. First, we have "a set of", because an act isn't considered an ability if it is only present under one circumstance, be it factual or counterfactual[43]. That's why we say there is a *set* of circumstances under which acts need to be present. This draws open the full range of circumstances in which S will (and won't) act, and allows us to characterise and assess the stability of an ability (namely as appropriate acts under a range or repetition of salient circumstances, regardless of whether these circumstances obtain - more on that in Section 2.1).

On to the "salient" part. We bring in counterfactuals only because we are interested in those that resemble obtainable or expectable factual circumstances. Not every counterfactual situation is illuminating for this, so not all counterfactuals are relevant for our purposes. If I want to know whether you are able to produce a proof for the irrationality of the square root of 2, then I will be interested in the various factual and counterfactual circumstances where you do and do not produce this proof. But some of these will be inconsequential or irrelevant. For instance, a counterfactual universe where the subject has had excessive mathematics training is inconsequential to the attributions of a subject who is in fact untrained, but a counterfactual where the subject is awake is not inconsequential to the abilities we attribute while she is factually asleep. Other counterfactuals are downright irrelevant. For instance, a counterfactual universe in which terrorizing mathematics-hating aliens live among us is not very relevant (even if it would be a difference-maker). I am here sailing close to Greco's (2000) characterisation of abilities:

[43] Technically, it would not be incorrect, since we can say "able under those precise circumstances". But everyday ascriptions of abilities certainly go beyond this.

> "[A]bilities in general are functions of success in relevantly close possible worlds. In other words, to say that someone has an ability to is to say that she would be successful in achieving X in a range of situations relatively similar to those in which she typically finds herself." (Greco, 2000, p. 13)

We want to exclude possible worlds that are too different or outlandish from the ones we may expect to obtain, so I fully endorse the addition of "relevantly close possible worlds". If the similarity of possible words is overly limited, we would only be considering identical worlds, and if it is too wide, it would deviate from the initial meaning, but somewhere in between works just fine. (Dennett & Taylor, 2002) But Greco also said "situations relatively similar to those in which she typically finds herself" and I'd prefer to leave the circumstances a bit more open by referring to them, simply as the "salient ones," similar to what Wilkenfeld (2013a) did earlier[44]. Presumably, the circumstances in which she typically finds herself will be salient, but typical circumstances needn't be salient and salient circumstances needn't be typical. If our subject could only work in a room which is colder than room temperature, then what makes her appropriate acts relevant is that those circumstances are salient, not that they are typical of the rooms she is in. In Chapter 2, I will offer up conceptual tools to determine which circumstances are salient. It would be beyond the purpose of this dissertation to give a specific theory of what makes a counterfactual appropriate, but I believe that for our present purposes our intuitions suffice to do this quite adequately without such a theory - at least for now.

Crucially, the assessment of a modal ability is not one where we literally peer into counterfactual worlds. Counterfactual worlds are not worlds we can discern with the naked eye.[45] They are merely useful conceptual tools for us to make claims (i.e. explanations or generalised predictions) about the circumstances under which we expect the acts to be present. Those claims are warranted by their explanatory or predictive power with regards to the subject's actual acts. As such, counterfactuals are more instrumental than metaphysical claims. The task of assessing understanding, then, becomes the task of identifying factual acts, and evidence of counterfactual ones. On the basis of these, we can derive generalised claims and predictions about the circumstances under which the appropriate acts would occur, or the defeaters that would prevent them (more on that in Section 2.3). This approach

[44] "(...) could (in counterfactuals salient in C)" (p. 1003/1004) - full quote in Section 1.2

[45] Counterfactuals have been criticised (e.g. Harris, 2012) for being scientifically untestable. However, I don't believe that is an entirely fair assessment, as science attempts to deal with counterfactuals all the time. If it didn't, the statement "were it thirty degrees Celsius, the snow would have melted" would be relegated to the realm of pseudoscience or metaphysics. Without the use of counterfactuals, we couldn't even say that a car can reach 50km/h unless that's the speed at which it is currently driving, yet no one is criticising car-salesmen for their outlandish metaphysical claims (or at least not for the 50km/h claim).

fits with everyday justifications of understanding-attributions. When asked to motivate whether a student is able to produce a proof, we are not relying solely on the student's actual production of that proof at the time of attribution. So, if a student has proved the proof in the past under numerous circumstances, but at this precise moment she seems overly nervous, then the claim "she could do it if you wouldn't stare" seems justified without our colleagues shouting "metaphysics!" at our assessment. Claims such as "she would have been able to do it if you hadn't been staring" translate into conditional counterfactuals claims such as "She, barring some significant changes[46], under circumstances that are saliently similar that do not involve your staring, would do it," which can be warranted by her (without significant changes, such as extra studying, years of changes, brain lesions or a new brain[47]), under similar circumstances that do not have anyone staring, producing the proof. In short, the use of counterfactuals is act-based, and the explanatory work they do is in grouping certain appropriate acts as to be expected under certain circumstances.

Meaning of an Object

On to the next concept that seemingly has nothing to do with acts, namely the *meaning* of an object. When someone is said to understand, then there is something, an object X, that is being understood. Object X is the *object of understanding*, what the understanding is *about*. There are several objects or types of objects that one could understand: a proof, theorem, problem, concept, strategy, (Sierpinska, 1994), theory, event, model, mechanism, causal relationship, state of affairs, etc. Depending on which object the understanding is about, we'll have a different set of abilities which will be salient to understanding it. This entails the characterisation of understanding should involve a focus on how the object of understanding guides our understanding attributions of a subject. Such a focus is often made independently of a subject, leaving out the pragmatic (what it takes for a subject to understand that object) entirely. But if objects are things that can be understood, the characterisation of objects needs to either directly or indirectly indicate what it takes for a subject to understand it. An umbrella term

[46] I should point out that notions like "abilities", and "similarity" are all of a higher ontological category. We are not defining abilities (e.g. proving), circumstances (e.g. staring) or subjects (e.g. Alexandria) as one specific configuration at the physical or atom-level. (Dennett, 2004) They are informal predicates which are multiply realizable. Although vague and subjective, they don't cause any unusual problems (Dennett & Taylor, 2002). So if we are talking about some circumstances, certain subjects or their particular acts "staying the same" or "similar," we aren't talking about exactly the same general or local state of the universe, but whether the salient informal predicates apply (e.g. Alexandria hasn't changed and she has produced the proof under similar circumstances) or make a relevant difference (e.g. the staring was the problem). More on this in Chapter 2.

[47] Counterfactuals can also help us consider difference-makers in potential, for instance: she would be able to do it when she, for example, studies a bit longer, lets it sink in for a week, or has a brain tumor removed. More on potential in Section 2.1.

(or synonym) for characterising the object, by itself, is via the "meaning" of that object.[48] When we say Lindsay understands global warming, the *object* of her understanding is global warming. So whatever captures the meaning of "global warming" will need to, directly or indirectly, indicate what it takes for Lindsay to understand it. It is generally agreed that this will be broader than just symbolic or linguistic understanding[49] (what it takes to understand a word or phrase), but not distinct from it. Let us now take a stab at how one can characterize the object X, such that we have a way to determine what it takes for subject S to understand X. I'll argue that all these meaning-determining candidates, to the extent that they have conceptual coherence and power, still boil down to abilities.

When asked to characterise an object and its meaning, one possible avenue of doing so would be to cite the relevant formal definitions and permissible moves, tactics or rules in the (science) game[50]. But does this adequately capture the meaning of the object? And how does this relate to a subject understanding the object? Reducing the meaning of an object to its formal definition and rules of the (science) game actually has a few problems. Firstly, they often fall short of exhausting the object's meaning. The clearest argument against this claim is in the field where formal definitions and rules seem to be the most crucial: mathematics.[51] Mathematics and its practices are standardly characterized as involving exclusively formal definitions and moves. But recently, this characterisation has been much criticised by philosophers of mathematical practice as a misleading distortion of how mathematics is actually practiced. (see Van Kerkhove & Van Bendegem, 2007; Mancosu, 2008) For instance: proofs, in practice, are not (usually) formally admissible deductive paths from axioms to theorem. Most proofs in practice are informal. There is some discussion whether an informal proof is an indicator of (e.g Azzouni, 2004) or recipes (Avigad, 2010) or outlines (Van Bendegem, 1989) for formal proofs or something else entirely (Macbeth, 2012). But it is agreed that formal rule-following does not exhaust mathematics.

> "One of the salient features of a logical deduction in the course of a proof is that the deduction depends on an understanding and on prior assimilation of the meanings of the concepts from which certain properties are to follow logically. It won't do to say that in practice this is just a matter of using the definitions of the concepts in the course of a

[48] This is why understanding is so often synonymous for "grasping the meaning". (e.g. by Dewey, 1933; Sierpinska, 1994) What it then takes for a subject to understand, is often completely shifted to what "grasping" entails, which isn't always illuminating, as we have seen in Section 1.2

[49] For discussions on linguistic understanding, see (Barber, 2003).

[50] Quine (1990) sees science as one of Wittgenstein's language games, with prediction as the checkpoints.

[51] I'll be talking about objects of understanding in mathematics, but it must be kept in mind that I use "object" as "object of understanding," not in its technical mathematical sense.

> proof; for we are back at the issue of grasping the meaning of the definition and of using it 'logically' on the basis of that understanding. Anybody who has taught mathematics knows that even in a graduate course or research seminar, writing on the board a formal definition without detailed explanations of the intended meaning is a sure way to block comprehension." (Rav, 1999, p. 29)

Secondly, we don't always have a definition, along with rules of conduct. The concept of a mathematical proof has been difficult to pin down (for a playful text that makes this point clear, see Davis & Hersh, 1998 p. 4-6), but no one is claiming that mathematicians don't understand the meaning of proof. One could still protest that there is an exhaustive and operative definition of proof, but that it is merely implicitly known or understood. That may be so, but there's no guarantee for this. So, not surprisingly, being able to cite the definitions and permissible rules of usage of an object also does not exhaust the abilities involved in understanding it. However, that's not to say that rules and definitions are irrelevant. What definitions and rules does tend to capture is constraints on what is or is not appropriate. As such, they do help us judge the abilities of a subject by considering whether that subject uses the object in concordance with what the definition and rules stipulate. Directly spelling out the full scope of usages connected to an object would be quite difficult, which is why the indirect route of providing definitions and rules are still helpful ways of indicating which are some of the appropriate usages.

Another popular approach in characterising the object of understanding and what it means (i.e. what it takes to understand the object) is through a set of *propositions*. A proposition is a thing (often a sentence-like entity, sometimes a representation) that bears a truth-value (i.e. it can be true or false) and which has a certain meaning. For example, we can characterise the object "global warming" as a set of propositions which includes:

- There is an increase of global temperatures
- Methane is a greenhouse gas
- Cows produce methane
- Humans breed vast quantities of cows
- CO_2 is a greenhouse gas,
- CO_2 is produced by fossil fuels
- A lot of transport relies on fossil fuels
- Greenhouse gases block heat from escaping

- Without heat escaping there is an increase of global temperatures
- There is an increase of greenhouse gases
- ...

If objects of understanding are propositional, then any object could be fully captured by a (potentially infinite) set of such propositions (with varying degrees of specificity). There is some controversy about whether we can reduce all objects of understanding (or knowledge) to a series of propositions (more on that in the next subsection). But even if everything about an object can be captured in propositions, we still have to clarify what these propositions are. The list I gave earlier is composed of sentences in English, but propositions are more than just a string of letters. Propositions have a content, a meaning. This is often seen as composed of intensions à la Carnap's extension-determiners: intension (meaning) determines extension (what it applies to or designates - what it is about), but not the other way around. So propositions are characterised by what they are about (i.e. if two propositions are about something else, they cannot be the same proposition), but not the other way around (i.e. even if two propositions are about the same thing, it can be in a different way[52]). (Dennett, 1990) This is giving us directions about how to distinguish propositions, but it does not reveal much about what a proposition is or how a subject is purported to capture its meaning. Unfortunately, there's no single stable view about this and many epistemologists will use the term proposition without specifying which account they are following.

If one is willing to take on a heavy ontology, then propositions could be characterised as abstract entities, out there, waiting to be judged and asserted. They are correlates of the Platonic properties and relations, existing independently of minds and spatio-temporal realm. (Jubien, 2001) Their meaning and truth-value is determined by this realm and understanding this meaning, then, would involve the subject having access to (or have its reasoning stand in some correspondence-relation with) one of these propositions.[53] Frege calls this special relation between the subject and a proposition one of "grasping a thought" (Hanks & Hanks, 2015, p. 3) Dennett (1990), among many others, criticizes Frege for not giving an explicit account of what work we may expect of this "relation":

[52] For example, claims about Clark Kent and the same claims about Superman don't express the same meaning, so aren't the same proposition.

[53] It is an appealing notion, especially for mathematical understanding. According to Penrose (1999), when we understand a piece of mathematics, we have access to a Platonic world of mathematics. It's clear to him to him that we can access this realm (and that machines can't), even though he never clarifies how we do.

> "What mysterious sort of transaction between the mind (or brain) and an abstract, Platonic object - the Thought - is this supposed to be? (...) This question invites an excursion into heavy-duty metaphysics and speculative psychology" (p. 123)

Churchland (1979), also points out how little revealing it is to invoke a "relation".

> "The idea that believing that p is a matter of standing in some appropriate relation to an abstract entity (the proposition that p) seems to me to have nothing more to recommend it than would have the parallel suggestion that weighing 5kg is at bottom a matter of standing in some suitable relation to an abstract entity (the number 5)" (Churchland, 1979, p. 105)

When it comes to specifying what is involved for a subject to understand or believe the proposition in question, the "relation"-move is not one of clarification, but one of obscurification. Hanks & Hanks (2015) give a more favourable reading of the Fregean view by denying there really is a special relation with a Platonic realm. Grasping, they say, is merely metaphorical. However, they do little to clarify an alternative beyond that what the subject does is "identifying" (p. 14) or "singling out" (p. 14) such a mind-independent entity by "comprehending, or understanding" (p. 15) them. This latter concept, according to them, philosophers would take to be primitive and not in need of clarification. Once again, we have a case of the synonyms, except this time we have come full circle.

In a more Wittgensteinian approach, I'd like to offer that the meaning of an object X corresponds to the several appropriate uses of (and surrounding) X as part of a(n institutionalised) practice. This would entail that grasping a meaning is displaying the appropriate usage. This usage is already implicitly hinted at in formal definitions and rules (where we specify a way to distinguish which acts are appropriate and which are not, even if it doesn't always give us a clear or exhaustive guide on when to use what), and propositions (where we specify a list of appropriate assertions and a rough guide on what these assertions entail). Furthermore, this characterisation side-lines a deeper metaphysical nature for these objects (as well as the need for a bridge that allows the subject to reach it). The "aboutness" of meaning here doesn't have to stand in relation to something that exists materially or in some Platonic concept world. As long as their use is semi-consistent and/or shared, these objects can be said to exist virtually (i.e. existing by virtue of being treated as if they exist). It also acknowledges that it is the practice (which employs these objects) that determines the truth, nature and relevance of what characterises these objects. The meaning is determined by the practice

to which the object in question belongs. Mathematical objects are, as Godino (1996) remarks, socially shared cultural entities (expressed in a symbolic language), progressively emerging, evolving and systematized via the human activity to solve problems and share the problem-situation and its solutions (i.e. a purposeful practice). Capturing or grasping this meaning, under this view, consists of possessing the abilities that exhaust (to some extent) the uses of that practice. Godino can thereby sum it up rather nicely, I think:

> "[A] subject (...) 'has grasped the meaning' of a concept, if the subject is able to carry out the different prototype practices that make up the meaning of the institutional object." (Godino, 1996, p. 6)

A downside of this approach is that it makes no sense to speak about understanding objects that are radically new to a practice, because there are no obvious candidates for appropriate usage. It may be fair to assume that, when a practice adds objects to its repertoire, the uses it decides are not arbitrary. But whatever makes them non-arbitrary is not something that can be accounted for here (outside of its relations to existing appropriate usages of existing objects within the practice). While this is a downside of my contextual account, I do not believe it is a big one since it also side-lines having to specify, with heavy normative import, how scientists or mathematicians should make these decisions, now and forever.[54] It is not my place to say, so I won't.

In sum, the work that is done by meaning-characterisation boils down to acts, and the explanatory work that they do is in indirectly signifying or grouping certain acts as appropriate. As long as one uses "meaning" in this sense, it is perfectly possible to talk about characterising the object through propositions, definitions, etc, even in an ability approach like my own, because it means that each of these ultimately translates to, or indicates at, appropriate usages. Furthermore, such characterisations don't need to exhaust the full meaning of the object to be useful[55], but on pain of vacuousness or obscurification, it is important that such claims do boil down to usage.

Mind of a Subject

The last concept that may seem orthogonal to the ability-approach is that of a mind. We've already established that we can't place our premium on marking understanding through what happens "inside," because "inside" is not a place we can (or need to) peer into. But this does not preclude us

[54] We'll come back to discussing some (broad strokes regarding the) contextual determinants in Section 2.2.

[55] If everything boils down to appropriate performances, then propositions can be an indirect way of signaling those, but there's no reason to assume the reverse: that all kinds of performances can be signaled by propositions.

from talking about "internal" concepts altogether. Our main lesson was that what demarcates understanding can't boil down to something more defining *behind* the performance which we can't (or simply don't) discern adequately - especially if it is the external acts (or evidence of their potential) that actually guide our ascriptions. But in a functionalist spirit, we can still instrumentally postulate what happens inside the brain or mind, as long as these postulates deliver explanatory or predictive power towards the subject's acts. The most ubiquitous "internal" postulate for understanding is that of "belief". When someone comes short of understanding, we may attribute this to a lack of the appropriate beliefs (e.g. Hilde doesn't realise it takes more plants to produce a hamburger than to produce a plant-based burger) or even false beliefs (e.g. Hilde believes livestock grows on trees). There's a whole array of philosophers who accept that understanding involves beliefs or has a belief condition (e.g. see e.g. Grimm, 2011; Khalifa, 2013; Kvanvig 2003; Hills 2009; Pritchard 2014). But it is far from clear whether the concept of belief accurately tracks something salient in the world. Beliefs "have a less secure position in a critical scientific ontology than, say, electrons or genes" (Dennett, 1990, p. 117). Do beliefs really exist? Arguing that they do is not as easy as it seems obvious. According to eliminativists (e.g. Churchland, 1992), they don't and beliefs are part of a pre-scientific folk-psychology theory that will one day be replaced by a superior scientific theory that tracks patterns in the world with more accuracy. (Schwitzgebel, 2019) So if we want to keep talking about belief, we must know what is that we are talking about. Not only should we have a way of pinpointing them and distinguishing one from another, but we must also have an idea of the kind of explanatory or predictive work beliefs are delivering. In short, we need to know what sort of concept "belief" is. While most contemporary epistemologists readily use the notion of belief, they don't usually clarify the explanatory work it is purported to deliver.

The standard concept of beliefs is as a propositional attitude, where a proposition is the object of the attitude, and the attitude is one of belief. The characteristics of propositions, as mentioned in the previous subsection, carries over here: they can be true or false (two beliefs are not the same if one is true and the other false), it must be composed of extension-determining intensions (two beliefs are not the same if they are about different things, although two beliefs can be about the same thing in different ways and not be the same belief), and they must be graspable by a mind. If it was unclear whether all objects of understanding could be exhaustively characterised by a set of propositions, it is now equally unclear whether a subject's understanding of that object could be exhaustively

characterised by propositional beliefs. Some epistemic abilities are not easily translated into propositions.[56] But even if they can, there are further problems to be overcome.

Propositional beliefs need to play a functional role in the subject's production of behaviour to be meaningful. Of course, if someone has a belief, this is a claim that cannot be substantiated by the mere utterance (e.g. a preschooler saying "Mummy is an orthodontist"). A belief more naturally expresses itself by the *several* uses that are made of it (of which the construction of a picture or the utterance of a sentence may be just one). There is a "tacit presumption of mental competence that underlies all belief attributions; unless you have an indefinitely extensible repertoire of ways to use your candidate belief (...) in different contexts, it is not a belief in any remotely recognizable sense." (Dennett, 2013, p. 66)

A common and more functionalist (but I believe still somewhat misguided) approach to flesh out a subject's attitude towards a proposition, is representationalist: a belief is some entity (e.g. an image, a sentence,...) being contained in the belief-box of the mind, such that it plays a causal (or functional) role in producing the subject's behaviour. Of course, the mere mental presence of a particular sentence (e.g. the sentence pops up) or any image thereof (e.g. a mental picture of mummy at work) would be equally insubstantial if it doesn't play a functional role in the production of the subject's behaviour. So a subject only believes or "grasps" a proposition if the subject has some concrete instantiation in her psychology which mimics the proposition (i.e. it acts as a representation). This means that something in the mind or brain of the subject is isomorphic or homomorphic to the proposition such that it plays the same role in the subject's psychology as it does as the content of a proposition. A famous proponent of this approach is Jerry Fodor. Fodor conceptualises belief as sentences in a language of thought parked in the belief-box of the mind (Schwitzgebel, 2019; Dennett, 2013). Analogous to machine language in computers, our brains have a language of thought, and for each belief that a subject has, we can find a sentence encoded in that language of thought. Those beliefs are mental states, and they have a structure that has the same syntactic and semantic content as the corresponding sentences (making them a representation of what is believed). (Schwitzgebel, 2019; Dennett, 1998) Unfortunately, these are strong and specific claims about the way the mind

[56] Consider a parallel in the literature on knowledge: Some knowledge is composed of (or includes) know-how, but whether know-how is propositional (a kind of know-that) or not is heavily disputed. According to intellectualists (e.g. Stanley & Williamson, 2001), we can reduce all know-how to know-that (i.e. propositional knowledge). According to anti-intellectualists (e.g. Ryle, 1949/2000), we cannot reduce know-how to know-that because know-how often stands or falls with abilities (or dispositions). (Fantl, 2017) This problem is even more palpable for understanding, where abilities play a more prevalent role (even if one only sees abilities as symptoms of mental states). Some kinds of understanding (e.g. some forms of causal understanding and social understanding) involve competences that aren't easily characterised by a list of propositions.

achieves the appropriate behaviour (namely that the mind represents the belief in a language of thought that has the same syntactic and semantic content as sentences) which furthermore lead to some conceptual problems as well as empirical ones. Much like an earlier cited problem with mental states, each belief would need to be encoded explicitly[57] and we would need to accept the assumption that what makes something a belief (and especially what makes two beliefs the same belief) has a specific kind of encoding (which is shared between two people with the same belief), which is an open question at best. It also fails to account for the fact that no one can have a single belief. Beliefs are inherently intertwined. You can't have the belief that a dog has four legs if you don't also have beliefs about what legs or dogs are. In another example, imagine a single belief "I have a sister, living in Cleveland" being added to a subject's belief-box. If a stranger asks her whether she has any siblings, what would happen? There's no belief about the sister's name, or about events that include her. In fact, there's absolutely no further indication of the belief other than perhaps a knee-jerk utterance of "I have a sister, living in Cleveland," which would be more like a tick than a belief – a tick which would furthermore confuse the subject as much as the stranger asking the question. (Dennett, 2013)

Another way to flesh out a propositional attitude is as a psychological predicate: a belief is some abstract propositional predicate useful in describing the subject's psychology in such a way that variations in psychology vary directly with propositional attitudes (i.e. changing one's propositional attitude entails a change in psychology, and people who share a propositional attitude, share a psychological state). Beliefs (as propositional attitudes) would be vindicated as a real predicate of a subject if there were a direct link between propositional attitude predicates and psychological predicates. But propositions as psychological states is a theory that faces counterexamples where the psychological predicate is the same, but the propositional attitude predicates are different, showing that psychological state and propositional attitudes come apart (Putnam's Twin Earth thought-experiment[58] is the most famous argument that meaning in belief-ascriptions is not wholly determined by psychological state). Furthermore, it is an empirical question whether it will indeed be so easy to link up psychological predicates with propositional ones. (Dennett, 1990)

Nevertheless, even when there is no direct link between beliefs and a person's psychology, that doesn't mean "beliefs" don't pick out anything salient. Dennett (1998b, 1990, 2009, 2013) argues that

[57] For an attack, see Dennett (1978a).

[58] Suppose there is a Twin Earth, an exact duplicate of Earth, except where we have H_2O, the duplicate has XYZ instead. On Earth, you have beliefs about water, but your duplicate has the corresponding (duplicate) beliefs about XYZ instead. Even though you are both physically (and thus psychologically) identical, what you mean with "water" (and thus with a water-based proposition) is not. So "meanings' just ain't in the head" (Putnam, 1975, p. 114)

beliefs are only as real as their explanatory or predictive power - and this may only be approximately (and sometimes ambivalently) so. To make this clear, he subsumes beliefs as one of the attributes projected onto entities as part of taking *an intentional stance* towards it. Here's a good summary of the intentional stance:

> "Anything that is usefully and voluminously predictable from the intentional stance is, by definition, an intentional system. The intentional stance is the strategy of interpreting the behavior of an entity (person, animal, artifact, whatever) by treating it as if it were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires.'" (Dennett, 2009, p. 339)

According to Dennett, the intentional stance is an innate capacity (as opposed to an academic theory) to interpret an entity as being governed by beliefs, intentions, and rationality. The sole justification for considering an entity as an agent (with beliefs) is the efficacy of the stance in predicting or explaining its behaviour that way (regardless of how it is realised physically), so there's no difference between a "real" and an "as if" agent, and no dividing line between the two. (Dennett, 2009) It is a normative stance in that this interpretation depends on what the agent ought to do (rationally), and requires a holistic approach in that the success of the stance lies not in pairing up the components of the stance with particular behaviours, but in how well the agent-package predicts or explains the entity's behaviours overall. (Dennett, 1990) If there is explanatory and predictive power to the intentional stance, then there must be some real pattern in the world that the stance exploits. But the intentional stance makes no dictates on what it is a pattern of, so it does not rely on there being a direct link between our ascriptions of belief and some structure in the brain. (Dennett, 1990, c5) In (1998, c22) he offers a good analogy of beliefs with dollars to show how they are both abstract concepts to track something salient in the world:

> "How many dollars (...) was a live goat worth in Ancient Athens? (...) [N]o one doubts that dollars are a perfectly general, systematic system for measuring economic value, but I do not suppose anyone would ask, after listening to two inconclusive rival proposals about how to fix the amount in dollars, "Yes, but how many dollars did it really cost back then?" There may be good grounds for preferring one rival set of auxiliary assumptions to another (intuitively, one that pegs ancient dollars to the price per ounce of gold then and now is of less interest than one that pegs ancient dollars to assumptions about "standard of living," the cost per year of feeding and clothing a family of four, etc.), but that does

> not imply that there must be some one translation scheme that "discovers the truth."" (Dennett, 1998c, p. 328-329)

So beliefs do track salient differences in the world, but they may track these differences only approximately and/or ambivalently so. This is a key difference between the intentional stance and the propositional attitude approach.

I will come back to the intentional stance extensively in Chapter 4 (and onwards), but for now it suffices to conclude that the even the concept of beliefs can be rooted in acts, and that the explanatory work they do is in explaining a behavioural profile with instrumental postulates. So even the beliefs we associate with or derive from attributions of understanding must yet again boil down to abilities, providing further validation to the ability account.

In Sum

Given that the value of understanding is hard to deny (since understanding is a valued aim and trait in many activities and disciplines), and that the value of its mark is no longer denied (since the concept of understanding has dissociated itself from its psychological dimension, as well as distinguished itself from the concepts of explanation and knowledge), we are in need of a conceptual characterisation that is explanatory as well as philosophically coherent and consistent.

In this first chapter, I focused on the *mark* of understanding, namely which systematic trait we find so philosophically or epistemically valuable about understanding and thus necessary for its attribution, regardless of who (i.e. which subject) it is attributed to or what the understanding is about (i.e. the object of understanding). I called this the “mark of understanding”, because it is what demarcates it. This mark of understanding needs a philosophically coherent and explanatory characterisation that can be applied consistently to various human subjects (and possibly beyond - which will be the focus of Chapters 4 through 6) and across various objects with varying degrees of (contextual) quality (which will be the focus of Chapter 2). Furthermore, it needs to allow us to deal with the known philosophical problems of marks, and address possible counter-examples (which will be the focus of Chapter 3). With inspiration drawn from Ryle (1949/2000), I have argued that understanding-attributions always boil down to a particular set of appropriate abilities (of a subject), composed of acts (salient to the object for a certain context), and that this is the most coherent and useful conceptualisation of “understanding”. Furthermore, I covered some of the useful concepts associated with understanding

that do not obviously match with an act-based approach and showed how they can not only keep their explanatory power, but are even more firmly rooted as instrumental concepts derived from acts.

I started with the benefits of the ability approach: we side-line, without discarding, the mistrusted role of feelings. We avoid some of the problems that plagued mental state-based approaches, such as locating its mark in an empirically unobservable realm (we cannot discern anyone else's mental states, and even struggle adequately characterising our own, but we can detect the acts of a subject), its explanatory redundancy (it is not the mental states themselves that are empirically accessible or epistemically valuable to us, so we both detect and judge mental states by the abilities, and not vice versa) and its requiring infinite encoding (every component of understanding would need to be encoded as a state). Furthermore, the concept of implicit understanding is given more room to flourish (because epistemic abilities can be valued even without the subject being characterised as "aware" of them), and the problem of implicit chauvinism is given *less* room to flourish (it is harder to substantiate that a particular gender, ethnicity or even species lacks understanding if one has to mark a valuable difference in performance rather than in physical or presumed mental constitution).

Ability-based approaches do entail considering what lies *beyond* observable acts (through explanatory estimations based on observed acts, conceptualised modally as counterfactual acts), but not what lies *behind* them (in an empirically unobservable realm), as mental state-based accounts presumed were necessary. The ability-based approach does not, however, preclude us from using "internal" concepts such as beliefs, provided they are instrumental postulates that function as an explanatory interpretation derived from the way the subject acts (as is the case in interpretationist approaches to the mind, such as the intentional stance - which will be further developed in Chapter 4).

Finally, I briefly considered some candidate kinds (or brands) of abilities offered by the literature as the appropriate one(s), to indicate that I will consider none of them as the necessary or sufficient condition for understanding, but instead as what *composes* understanding. This will allow the quality of understanding to be expressed through the amount of salient abilities (which I be further develop in Chapter 2) and will allow understanding to vary its salient abilities with the meaning of the object that is being understood (which we can conceptualise as the appropriate usages or indications thereof) for each context of attribution (which I will also develop further in Chapter 2).

PRELUDE 2

A Suitable Suit

A voice can be heard. It's yelling "Stop! Stop! Stop!"

STOPPARD: Stop! Stop! Stop! Give o'er the play. You're writing straw-men. These students are made of straw, they are straw-students!

SVEN: I know, I'm sorry.

STOPPARD: Then why are you doing it?

SVEN: Three reasons. The first reason is because I want to make an intuitive version of my point in a funny way, which helps the reader follow my trail in a way that avoids the dry academic writing.

STOPPARD: But it's still a bit of a disingenuous approach, is it not?

SVEN: I know it is a bit, but that's why we're addressing it right now (and why there is also the academic writing). To see the full scope of the *picture*, we have to *draw* open every aspect.

STOPPARD: What a silly pun. But let's not let jokes side-track us. The problem with the *That Within Which Passeth Show* dialogue, or, as I like to call it, the *Rosencrantz and Guildenstern Are Dead Wrong* dialogue, is not that abilities should really have been equated with understanding, but that such situations (where someone has understanding without abilities or vice versa) would never happen to that degree. Understanding will always be *accompanied* by abilities. The appropriate mental models always align with abilities, to varying degrees. That's why they're the *trappings* of understanding. But they can also mislead. That's why they're also the *traps* of understanding. No one would deny that people can have trouble displaying their understanding, not even you.

SVEN: Indeed I don't. But what do you gain from conceptualising understanding as something *behind* the abilities if it is entirely validated by the abilities? Why not save yourself the trouble and conceptualise understanding via the abilities directly? Why must the suit of understanding be its trapping and not its feature?

STOPPARD: Because sometimes students display abilities, even though they don't understand. For instance because they've just memorised the responses or copied those of their neighbour's.

SVEN: So why don't they understand? What makes you so sure of this?

STOPPARD: All they can do is repeat the memorisation, or endorse their neighbour's answer. Do you call this understanding? I thought you'd be a better philosopher than that. Isn't it clear that, in such examples, they don't grasp a mental model when they give these answers?

SVEN: What is of interest to me is that even while you state their lack of abilities (namely doing anything *beyond* repeating an answer), you still feel the need to conceptualise the

problem as something lacking *behind* it. Isn't it once again the lack of apparel, or suit, that proclaims the man? What does this extra move accomplish except a redundant speculative metaphysics which leads to misleading assessments exactly like the ones from the first dialogue?

STOPPARD: But sometimes people don't display abilities, even though they do understand. They have that within which passeth show. For instance, did you know Hamlet has all these beautiful and elaborate speeches?

SVEN: They're certainly elaborate. But most of it is pompous jokes, old-fashioned pandering and outdated references that are no longer appropriate for contemporary audiences.

STOPPARD: Fair enough, but the point is he does have these speeches, even if he doesn't show them to anyone.

SVEN: But how do you know he does? Is it because you were able to hear into his mind's mouth, or because you were a secret witness to his performed soliloquies? You can't set up a glass where you see (or hear) the inmost part of him, whatever that is actually supposed to mean. If he *never*, under any circumstance, puts on his speech-cap, no matter the opportunity or pressure, then we may have to simply conclude that the Prince has no clothes.

STOPPARD: Pragmatism, is that all you have to offer?

SVEN: What more can you offer?

STOPPARD: You're doing it again. You're making me into a straw-man. And me, a writer you admire, no less!

SVEN: In my defense, if it had been inappropriate for the dissertation, I wouldn't have done it.

STOPPARD: Speaking of, you never gave me your third reason for these dialogues.

SVEN: Quite simply because I think dialogues make for a better text. I appreciate ideas being dressed up in a more playful suit.

STOPPARD: That depends what the text is written for. There's nothing either good or bad, but context makes it so. And for a dissertation, I'm not sure it is appropriate since it's not academic writing. This playful suit doesn't quite suit it.

SVEN: That'll be up to the jury to decide.

STOPPARD: Relativism, is that all you have to offer?

SVEN: Not quite, I can still make arguments why the jury's context can value - or, at least, not devalue - the role of dialogues in a dissertation.

STOPPARD: It's not difficult to see that dialogue-writing is not an ability required for academia, so showcasing it still seems inappropriate.

SVEN: It's not required, but it can be valuable even outside of meeting a requirement. Dialogues may not directly put forward my academic arguments, but they can indirectly evidence that I have understood them adequately. Therefore, I think the context of assessment should be allowed to include assessments of dialogues, even if they aren't judged for their artistic merit, and even if they are not nearly as important as the academic text in between.

STOPPARD: That'll be up to the jury to decide.

Chapter 2

ON EXPRESSING THE QUALITY OF UNDERSTANDING

Understanding often comes in degrees, meaning we need a way to both express and evaluate its quality. In this chapter, I will conceptualise the dimensions and degrees of quality in understanding, offer up a contextual approach to specifying what is salient, and specify some of the problems and opportunities in evaluating understanding under my approach. This will enrich my account enough to address many of the examples and objections (raised in the literature) that I will discuss in Chapter 3.

In this chapter, I will argue that the quality of understanding doesn't have to slide along a single axis. To begin, I will present four dimensions where a higher degree would lead to a superior understanding. The first is the *scope* of abilities, which tracks the amount of different abilities. The other three dimensions focus on the degrees of quality within each of these different abilities, and will consist of two parameters, one which widens it and another which deepens it. These dimensions will express how *sensitive* an ability is to the demands of a practice (comprised of the situational responsiveness and accuracy parameters), how *stable* the acts that compose it are across circumstances (comprised of its range and robustness), and how *efficient* the subject is producing them (comprised of the economy and potential parameters). It is my contention that most attributions of understanding will boil down to a claim about the degree within these dimensions (examples aplenty in Chapter 3).

Unfortunately, and quite unsurprisingly, no single agreed upon universal standard can clarify all attributions of understanding within these dimensions. Therefore, the next section will offer up a contextual approach to each of the dimensions and parameters. I will conceptualise how to express the contextual variations in each dimension (and each parameter specifically) by allowing the context of attribution to give more or less weight to the salience of specific kinds of abilities, circumstances or efficiencies, along with the option for thresholds. Additionally, I'll briefly touch upon the virtues of a contextual approach, as well as some of the determinants in what is to be considered a fair or scientific context of attribution (while leaving the full justification of what is appropriate to another discussion).

Lastly, I'll conceptualise some of the practical concerns in evaluating these dimensions and parameters by addressing some of its aspects. I will label a common misevaluation, introduce a distinction between direct and indirect evidence, point to some limits of characterising a context of attribution, show how contexts of attribution can also handle kinds of understanding, and consider what it might mean to have complete understanding.

2.1 Dimensions & Degrees of Quality

In Chapter 1, I mentioned that the quality of understanding can be expressed through the amount of salient abilities, and I employed terms that were intended to indicate at that amount (i.e. scope, sensitivity, stability), but I did not yet clarify how to conceptualise that quality or interpret those terms. This section is an attempt to conceptualise the dimensions, parameters and degrees of quality of understanding in a fruitful way. I will say upfront that I am far from sure that these dimensions and parameters are the best possible way to conceptualise the quality of understanding. They are not rigidly defined and the difference between them can sometimes be debatable (i.e. there's a potential overlap in what they capture). But even if they are imperfect in conceptualising the "ideal" assessments of the quality of understanding, they are fruitful in diagnosing the strengths and weaknesses in quality as well as the problems in evaluation - as will be attested by how well they fare in addressing tricky or misleading attributions and proposed counterexamples to the ability account (the topic of Chapter 3). Therefore, I believe they are a step in the right direction.

On Degrees & Dimensions versus Conditions

The language we use to describe understanding can mark it out as an all-or-nothing property. The answer to "does subject S understand?" is often "yes" or "no". And yet, understanding also clearly allows for gradation. The answer to "do you understand?" can equally be "a little," or "mostly". You can gain a "better understanding". One person can understand "better than" another.[59] This as opposed to knowledge, where the answer to "do you know that X?" is "yes" or "no", and not "a little". Better knowledge usually means more knowledge, and not knowledge of a higher quality. It is therefore generally presumed that you know something if and only if certain conditions are satisfied (e.g. that you believe it, that it is true, justified, and that you weren't lucky or gettiered). Under some accounts, understanding gets the same conditional treatment, and they compose it through a combination of conditions like a factivity condition, a justification condition, a belief or grasping condition, as well as a coherence condition, an anti-epistemic luck condition, etc. Many accounts rely heavily on a condition-centered approach, but is this the right approach? Are there necessary or sufficient (set of) conditions that, once satisfied, warrant the subject as having understanding? Van Camp (2014) favours an approach that can conceptualise understanding in terms of degrees, rather than through meeting a set of conditions (or even a threshold). I agree with this move. I will argue that if the presence of understanding depends on satisfying certain conditions, we will either be too strict

[59] See also (Hills, 2015) for further examples from language use.

or too forgiving and we won't account for the degrees of understanding. Understanding, as opposed to knowledge, requires an expression of not just its presence, but its quality. [60]

In my account of understanding, the focus is not on satisfying a particular set of necessary and/or sufficient conditions that mark the presence or absence of understanding, but on its dimensions and degrees of quality. There are a whole range of acts that can mark an ability and a whole scope of abilities that can mark understanding, but there is no exhaustive set of specifiable acts that are necessary or sufficient for an ability attribution and there is no exhaustive set of specifiable abilities that are necessary or sufficient for an understanding attribution. One of the problems with a condition-centred approach, as we'll see in this section (and Chapter 3), is that the proposed conditions tend to be too stringent or too loose, if not both. But once we let go of attributing understanding only if a certain (set of) condition(s) are satisfied, we can consider the gist those proposed conditions - not as necessary or sufficient conditions that need to be fully satisfied, but as multiple parameters which compose degrees of the dimensions of understanding. Then none of them needs to be singularly sufficient or necessary, but all can be relevant. Expressing the quality of understanding does not entail that we need to be able to quantify understanding on a numeric scale. But we do need a way to conceptualise the difference in degrees of quality.

Additionally, the quality of understanding doesn't have to slide along a single axis. I will present several dimensions where a higher degree would lead to a superior understanding. It is my contention that most attributions of understanding will boil down to a claim about the degree within these dimensions (examples aplenty in Chapter 3). We will cover four dimensions. The first is the *scope* of abilities, which tracks the amount of different abilities. The other three dimensions focus on the degrees of quality within each of these different abilities, namely how sensitive an ability is to the demands of a practice, how stable the acts that compose it are across circumstances, and how efficient the subject is producing them. Each of these dimensions (except scope) has two parameters (one which widens it and one which deepens it).

These dimensions and parameters are conceptual tools for expressing the quality of understanding. I do not claim that all understanding claims can be *best* explained by these proposed dimensions, merely that these are fruitful in doing so. The dimensions proposed are not meant to end the discussion, but to further it. And as I will showcase in the examples of Chapter 3, these dimensions

[60] It's worth bearing in mind that even if understanding is a species of knowledge, the all-or-nothing condition criteria does not necessarily carry over. If understanding is "more knowledge," then conditions need to be satisfied to mark the presence of knowledge, but it is the amount of things known that mark the degree of understanding.

and parameters do bear a lot of fruit. Firstly, they allow us to express a quality of understanding. Secondly, they allow us to pinpoint different kinds of understanding depending on how the subject fares in each parameter. Thirdly, they dissolve the need for certain conditions which have proven unwieldy. For instance, anti-luck conditions can be replaced by degrees (and a threshold) on the dimensions of understanding. Fourthly, they provide a basis for contextual differences in understanding attributions. More on that in the next Section 2.2, where we'll see how different contexts of attribution can place their values in different places in a systematic way, building on these parameters. And lastly, my account can easily incorporate something akin to all-or-nothing attributions by using a threshold. So binary attributions of understanding can be made by assessing a subject's abilities against some contextual threshold. So if each (or any) dimension gets a predetermined threshold, my account approximates a condition-satisfying one - except that it doesn't presume such thresholds to be necessary, and it doesn't deny that degrees are still at play. I'll come back to the notion of degrees after I've set up the dimensions and their parameters.

Scope of Abilities

I mentioned before that characterising understanding through abilities, plural, was not incidental. Understanding involves more than just a single ability (or even a narrow set of abilities). So if we want to adequately assess someone's understanding, several appropriate abilities need to be displayed (or inferred). Earlier, we also saw that there are multiple proposed candidates for the appropriate abilities for understanding (independently of whether the proposing author takes the candidate(s) as a mark of understanding or as symptom), including: recognising qualitatively characteristic consequences (de Reg & Dieks, 2005), making counterfactual inferences in the contexts of manipulation, prediction and explanation, give explanations (i.e. answering explanation-seeking questions) (Ylikoski, 2009), answering what-if-things-had-been-different questions (Woodward, 2003; Grimm, 2014), being able to evaluate explanations (Khalifa, 2013), relating knowledge to other knowledge (Van Camp, 2014), controlling the phenomenon (Yliksoki, 2009), specifying causal dependence (Ylikoski, 2014), reliably tracking dependency relations (Grimm, 2016), following an explanation given by someone else, explaining it in your own words, drawing the appropriate conclusion based on a given set-up, or vice versa (Hills, 2009), responding to challenges as to the correctness of a proof (or theory, or label), identifying key features, identifying the nature of the objects and questions, mustering the relevant background knowledge, exploring the space of possibilities fruitfully, and so on. (Avigad, 2008).

All of these are good candidates for which abilities are appropriate, but none of them are quite enough to fully mark understanding by themselves. The display of each of them makes an understanding

attribution more warranted, but understanding attributions do not stand or fall with any one of them. My contention is that understanding is constituted by them, but not exhausted by any one of them. But if understanding constitutes many abilities, then the amount of abilities present will improve the quality of understanding. The amount of abilities involved can be labelled as the *scope* of understanding.

When I discussed the object of understanding in Section 1.4, I noted that their meaning is to be found in the various appropriate uses connected to them (for a particular practice). To understand something, the subject would need to display its appropriate uses in some way. It is important to note that a particular object (as part of a particular practice) rarely, if ever, has just one single use connected to it. This entails that understanding X will involve more than one ability. From this, it is reasonable to suppose that the larger the set of uses that are covered is, the better the understanding will be. This may seem somewhat of an obvious point to make, but many of the proposed counter examples of ability-without-understanding (e.g. memorisation, luck, using a formula) fail exactly because of the narrow view they take on the abilities involved in understanding (see nearly all of the objections addressed in Section 3.2). Consider this quote from Ryle (1949/2000) about knowing how to tie a clove-hitch:

> "You exercise your knowledge of how to tie a clove-hitch not only in acts of tying clove-hitches and in correcting your mistakes, but also in imagining tying them correctly, in instructing pupils, in criticizing the incorrect or clumsy movements and applauding the correct movements that they make, in inferring from a faulty result to the error which produced it, in predicting the outcomes of observed lapses, and so on indefinitely." (Ryle 1949/2000, p. 54)

This is a claim about the *scope* of knowing how, but the same applies to understanding. When it comes to understanding a theorem, for instance, the aims and uses of mathematical practice are not exhausted by finding, owning or producing a formal proof for it. Mathematicians do prove, but they also reprove theorems (Dawson, 2006) and sometimes they do so often (Macbeth, 2012). This alone should be an indication that something else is valuable about proving other than the mere possession of a single proof. Furthermore, not all proofs are equally valued. Some proofs are considered better than others, because they, for instance, provide a better understanding. (Avigad, 2010) Proofs are not of interest solely as a means of verification, but in that they show key ideas, the fruitfulness of a methodology or concept, alternative routes, etc. (Lakatos, 1976; Rav, 1999; Dawson, 2006)

> "The whole arsenal of mathematical methodologies, concepts, strategies and techniques for solving problems, the establishment of interconnections between theories, the systematisation of results—the entire mathematical know-how is embedded in proofs" (Rav, 1999, p. 20)

If this is true, then it is clear why producing a proof for a theorem does not exhaust the abilities required for understanding it. Other abilities must be involved. You display your understanding of the theorem not just by supplying a proof, but in giving a rough outline of the proof, supplying different proofs, using the methodologies correctly, correcting mistakes in a faulty proof, using the theorem where it is appropriate to do so, explaining the concepts involved, showing what would happen if the theorem were false, etc. This does not entail that someone who understands must display *all* of the appropriate abilities, as well as all the abilities related to those (and abilities related to those abilities, and so on). But the scope, and therefore the quality, of understanding strengthens with the amount of abilities involved.

So much for the plurality of abilities. But when it comes to the quality of understanding, more is to be said about the quality of ability, *singular*. Therefore, the subsequent parameters will focus on the quality of a singular ability. With each ability, there are three further dimensions, namely how *stable* the ability is, how *sensitive* it is to the demands of a practice, and how *efficient* the subject is displaying these abilities. We'll first start with sensitivity.

Sensitivity of an Ability

In constituting an ability, the appropriateness of an act is itself subject to degrees of quality. Even when a subject acts appropriately, not all appropriate acts are equally fitting or precise. Abilities should be appropriately *sensitive*, be it to the situation or to detail. Sensitivity is comprised of two parameters, namely (i) *situational responsiveness*, and (ii) *accuracy*. I'll elaborate on each.

(i) Situational responsiveness

While discussing scope, I said that understanding is not usually exhausted by a single type of ability. But even a single type of ability does not usually get exhausted by a single type of act. Where scope was about the variety of appropriate abilities that can make up understanding, situational responsiveness is about the variety of appropriate acts that can make up one ability. The distinction between situational responsiveness and scope is mostly for practical concerns (as opposed to purely

metaphysical ones), so they may differ depending on the practicalities. One could, in theory, conceptualise all types of situational responsiveness as a different ability and thus subsume sensitivity under scope (or vice versa).[61] But, as I hope will become clear, distinguishing them can be fruitful in discussing the abilities involved (as well as their salience). If different acts fall under the same ability, then they may also add to that ability's *situational responsiveness*, the first sensitivity parameter:

> *Situational responsiveness:* the acts vary appropriately with variations in object-situations.

The first sensitivity-parameter is roughly similar to the amount of what-if-things-had-been-different questions that the subject can respond appropriately to. For instance: if we add a distant planet to our model, will the subject's predictions of planetary motion vary appropriately? If the concentration of greenhouse gases were changed, would the subject's estimation of the climate crisis change with it? If we alter an axiom, will it affect the proof that the subject produces? According to Hills (2015), this is the essence of understanding-why.

> "After all, how do you test whether someone really understands why global warming is occurring, or why stealing is morally wrong? You ask them a series of "What if...?" questions. What if the initial conditions were different? What would be the consequences? What if there was a different outcome? How could that be explained? If someone cannot answer these questions, they do not understand why p very well, whatever else they can do." (Hills, 2015, p. 11)

Being able to respond to these variations entails being responsive to object-situational variations. As such, it expresses the subject's ability to be sensitive to a(n object) situation. I will be using the word "situation" to refer to objectual variations, as opposed to "circumstance" which are supposed to refer to environmental variations. If a subject needs to solve an equation, then a change in the variables of the equation is a change in (object) situation, whereas a change in room-temperature is a change in (environmental) circumstances. It is, of course, true that situational differences can be expressed as circumstance-differences (where the environment presents a variation relevant to the object), but

[61] How many types of acts comprise an ability will depend, in each case, on how wide a net one wishes to cast with either concept. The ability "to multiply two single digit numbers" clearly casts a wider net than the ability "to multiply 5 with 6". As was mentioned in a footnote of section 1.3, I will leave open the size of the nets. This will not lead to any difference for my conceptualisation, except that the wider the net is of abilities, the smaller that of sensitivity will be and vice versa.

even so, they are a specific and very relevant type of difference in circumstance, so it is fruitful to distinguish the variation in affairs that are purely environmental from those that are objectual.

So situational responsiveness is a degree of principled variance of performance among varying object-situations. This makes it modal in nature. Modal proposals similar to this one appear in the epistemological literature in a variety of ways (both for knowledge and for scientific understanding). For scientific understanding, they are conceptualised as answers to what-if-things-had-been-different questions (Woodward, 2003) or the ability to make counterfactual inferences (Ylikoski, 2009) or counterfactual reasoning (Grimm, 2016)[62]. For knowledge, Pritchard (2008) talks about a safety and sensitivity *condition* - which are both combined here as a sensitivity *parameter* (so mind the difference in this paragraph between Pritchard's sensitivity condition and my sensitivity parameter). If subject S knows p, then Pritchard's sensitivity condition requires the knower to not believe p under circumstances where p is false, and the safety condition requires S to believe p under circumstances where p is true. As a condition for knowledge, some everyday attributions of knowledge would unfortunately get undercut by the sensitivity-condition (e.g. we can't know rubbish slides through the rubbish chute unless we can also tell when it is stuck) as well as the safety-condition (e.g. we can't know that our lottery ticket will lose unless we can also tell when we have the winning ticket). As conditions they are thus too strict, but as a parameter that sort of combines the two, it behaves exactly the way we want - not as a condition for understanding, but as a degree of quality in understanding. The greater the situational responsiveness, the greater the understanding (all else being equal).

(ii) Accuracy

The second sensitivity-parameter is intended to reward the degrees of precision or error in acts that can roughly be characterised as "appropriate". This is a fairly straightforward and well-known parameter of quality. You can distinguish between acts that are *more* or *less* accurate.

Accuracy: the ability has a degree of accuracy or precision (where possible).

There are two ways in which the concept of accuracy is of help. Firstly in expressing that the predicate "appropriate" does not constitute a dichotomy - not all acts are either appropriate or inappropriate, with nothing in between – for example, most calculations with rational or irrational numbers, predictions of planetary motions, or predictions of the effects of climate change in Belgium. What all

[62] To avoid confusion, I prefer to use the word "counterfactual" for expressing differences in physical circumstances and not object-situations.

these examples have in common is that they are, in some way, approximations with an allowable margin of error. And yet they are also examples where it is valuable to attain a higher degree of appropriateness. Accuracy is merely a way to help us express that degree of appropriateness. The greater the accuracy, the greater the understanding (all else being equal).

There is a second way in which accuracy is of help, namely in distinguishing an educated guess from an arbitrary one (or in distinguishing a reasonable mistake from a deranged one). Predicting that the earth will make a circular motion around the sun is incorrect, but it is not as incorrect as predicting it will revolve around the sun in the shape of a T-rex (or the shape of blue divided by apple). Having a parameter that allows us to express this difference is therefore fruitful.

Together, the situational responsiveness and accuracy parameters cover the sensitivity dimension, because it expresses how *sensitive* the ability is.

Stability of an Act

The scope and sensitivity dimensions focused on appropriate variations that were closely related to the object of understanding. Now, it is time to look at the dimensions which determine what's "appropriate" without necessarily having an object in mind. Abilities are comprised of several appropriate acts, but one displayed act does not an ability make.

> "Suppose the novice trampolinist's new coach asks [her] which tricks [she] is already able to do. The correct answer would not be a massive list including every trick [she] could pull off given some incredible stroke of luck." (Glick 2012, p. 129)

This entails that to have the ability to do something, it does not suffice to have acted appropriately under one precise set of circumstances. One must be able to perform it under various different circumstances. The larger the set of circumstances under which the appropriate act can be performed, the more stable (and thus better) the ability. This stability concept allows us to express a dimension of quality of the ability and, with it, also one of the dimensions of quality for understanding. An ability is *stable* if the appropriate acts (that comprise it) are present across and throughout circumstances. Therefore, stability will be comprised of two parameters: (i) *range* and (ii) *robustness*. I'll elaborate on each.

(i) Range of an Act

Part of what we described as stability is that one could perform the appropriate act even if the circumstances were different. Therefore, the first stability-parameter I'd like to suggest is one which is able to express the extent of that. To do so, we need to keep the subject constant and consider its performance under a range of circumstances.

> *Range:* The act occurs under a certain amount of salient (counter)factual circumstances, while keeping the subject constant (barring accessible non-epistemic changes).

If you take the same subject who is about to perform, place her under various types of (counterfactual) circumstances and it has little effect on the performance, then that subject has a wide range. For instance: the ability to produce a proof is stable if it can be carried out by our same subject regardless of variations in the circumstances under which she finds herself in (e.g. weather, time of day or location). By this I don't of course mean that she needs to do so under *all* circumstances, but in as many salient circumstances as possible (more on salience in the Section 2.2).

A small note is in order here. I said to keep the subject constant, so why did I allow "accessible non-epistemic changes"? The reason for this is quite simple: Otherwise, a subject who is asleep would have zero range, because if she's kept constant, she (being kept constant) would be asleep in all circumstances. Which changes are allowed? Those changes, firstly, which do not contribute anything epistemically salient (e.g. waking up is fine[63], years of study is not), and secondly which we may reasonably expect or can feasibly bring about (e.g. waiting a bit for the subject to lose grogginess is fine, but a complete brain rewiring is not).[64] I expect we could find contestable grey areas, but for most situations, our intuitions will clearly guide us in deciding whether a change is saliently epistemic or not. So, only if the change is not saliently epistemic, is what is relevant about the subject held constant[65] and only if the change is reasonable to expect or feasible to bring about will it usually be considered salient (see Section 2.2).

[63] It's true that sometimes you need to sleep on something before you start understanding, but if that is the case, we probably have indications here that the subject's range was absent before she went to sleep and wide when she wakes up. But that's not what was presupposed here. Here, sleeping merely *masked* the ability that was present both before going to sleep and after waking up. More on masks in Section 3.1.

[64] A more precise and contextual way to deal with "accessible non-epistemic changes" is to lump certain changes as part of contextually permissible growing potential (a parameter which we'll get to in the next subsection on system efficiency).

[65] In short, what we really need to keep constant is not so much the subject, as it is the epistemic subject - meaning everything about the subject that is epistemically relevant (more on the epistemic subject in Chapter 4)

Barring those inaccessible epistemic changes, the greater the amount of circumstances under which the subject (held constant) can perform appropriately, the greater the range and therefore the greater the understanding (all else being equal). As we'll see (in Section 3.2), requiring a certain degree of range already keeps many forms of lucky acts at bay.

(ii) Robustness of an Act

Now, a subject could, at a certain point in time, be prepared to construct a proof of a particular mathematical theorem no matter what the circumstances are (i.e. have a wide range), but still fail at retaining (and thus repeating) that performance for various reasons. When we gauged range, we needed to keep the subject (epistemically) constant. In gauging robustness, we will estimate how (epistemically) constant a subject would remain even after having gone *through* various circumstances. A subject's understanding wouldn't be of much quality if a little time or a few events (e.g. being told to think of something else) made her lose that competence. Understanding should be made of sterner stuff.

> *Robustness*: The act can be repeated even after the subject has gone through the salient (counter)factual circumstances

If you take a subject who is able to perform appropriate acts and let her go through various types of circumstances, and can determine that it has little effect on the appropriateness of her future acts, then that subject has a strong robustness. For instance: the ability to produce a proof is robust if it can be carried out by our same subject after going through various circumstances such as performing other tasks, learning about different things or hearing deceptive suggestions about incorrect proofs. A good ability sticks. Someone with a poor long term memory has little robustness, because any sequence of events that lasted long enough would have her lose all the relevant information to act. Once again, this does not mean that she needs to retain her acts no matter what happens to her, but it is important for her ability's stability that she retains it after going through as many salient circumstances as possible (more on that salience in Section 3.2). The greater the robustness, the greater the understanding (all else being equal).

Like range, the conceptualisation of robustness, too, involves reference to counterfactuals. We aren't just interested in how well a subject retains her ability through her actual future, but also through a set of counterfactual circumstances which she *could* go through. This is relevant both because we don't know the actual future and because we make attempts at steering the future based on what

could be expected. For instance: it is usually salient that a subject should not lose her ability to construct a correct proof as soon as she has been presented with an incorrect proof (e.g. being too easily swayed to copy anything that most recently looked like a proof), even if that subject might never actually run into someone who presents such a proof.

These were the range and robustness parameters. An ability is stable if the appropriate acts are present across (for range) and through (for robustness) the salient (counter)factual circumstances. Together they comprise the stability-dimension of an ability, and thus a dimension of quality in a subject's understanding. In Section 2.2, we shall reconsider these parameters (as well as others) in a contextual light to allow a practice to focus on which particular circumstances it deems salient. But for now, we move on to the fourth and last dimension I'll be discussing.

System Efficiency in a Subject

The last dimension is called system efficiency. This dimension is not often considered even though it is obviously valued. It roughly captures the following claim: The more *efficient* the system is that produces the acts, the more valuable it is. System efficiency is comprised of (i) *economy of resources* and (ii) *growing potential*. I'll elaborate on each.

(i) Economy

While separating a subject from external resources can be telling about her abilities, scientists (even mathematicians) don't usually work in a vacuum, so the evaluation of their performance should be able to include all of the used resources. That's why this first parameter expresses the value of using as *few* resources as possible. Resources include such things as time, tools, amount of brain(s), a library of information, a computer, a sheet of rules, or the amount of energy or food to keep the system running. Of course, not all of these resources are considered with the same degree of salience, but we'll address this in the next section when we consider the context of attribution. But if any of it *can* be considered salient, we need a way of addressing that, and that is what this parameter is for:

Economy: The appropriate act uses a minimum of saliently allowable resources

This parameter can be used as a way of delineating or demarcating the physical system that produces the relevant acts which make us attribute understanding. We'll consider the demarcation of a subject beyond the skin or skull more closely in Chapter 4, but even if we keep our subject neatly demarcated by skin or skull, the amount of *additional* resources needed to display the appropriate ability makes a

relevant difference to how we perceive the subject's understanding. If Charlotte and Olive are epistemic equals in all of the respects previously mentioned (scope, sensitivity, stability), but Charlotte is slow and Olive quick (i.e. the difference being the resource of time), then Olive has better understanding than Charlotte. The same is true if Olive only uses her own wits, and Charlotte needs to use pen and paper (i.e. the difference here is the resource of pen and paper). It is certainly true that a lot of these considerations often aren't relevant, but they are not always irrelevant either. Especially in a world that increasingly offers more technological resources. So we need to be able to distinguish the quality of understanding based on such resources. The economy parameter allows us to value people with quick and ready abilities, and this without discrediting the understanding of people who, for instance, need a little more time to respond. The fewer resources are used, the more economical the understander and thus the better the understanding (all else being equal).

From Chapter 4 onwards, I'll conceptualise the epistemic subject as an epistemic agent, a virtual postulate composed of beliefs, aims and rationality. This may offer up further possibilities of conceptualising the concept of economy. From the vantage point of epistemic subjects as epistemic agents, resource economy could also be considered from the point of view of virtual resources, namely how many beliefs does it take for the subject to display the relevant abilities? The lower the amount of beliefs necessary for the appropriate ability, the more economical the understanding, and therefore the better it can be deemed.[66]

(ii) Potential

The second system efficiency parameter is also about resources (in the broad sense), but not those which did bring about the appropriate act(s), but those which *would* bring about the appropriate acts. This conceptualises a subject's *potential:*

> *Potential:* The appropriate act obtains with the addition of a minimum of resources or events

Examples of resources or events include training, studying, a textbook, empirical research, or brain surgery. The concept of potential is about the amount of resources which make the difference in bringing about the appropriate act(s). It is a parameter that doesn't express abilities, but potential

[66] Other approaches to the epistemic subject will be mentioned in Chapter 4, and each may offer further opportunities to reconsider the concept of economy. For instance, from the vantage point of the epistemic subject as an information processing system it can be conceptualised as an economy of information. These are all possible routes for extending the current economy parameter.

abilities. As such, it is not about actual understanding, but about potential understanding. I'd like to add this parameter because it does speak to a person's credit of understanding if very little is missing (e.g. all that is missing is a piece of information, or a calculator, or some exercise) as opposed to when a lot is missing (e.g. a person would need to go through several years of training or brain surgery). If a subject needs to be reminded of the relevant facts, or re-discover some of the principles involved, then that doesn't entirely discredit the subject's understanding before she embarked on this process of recall or re-discovery. In fact, it is unclear whether understanding is ever really free of these types of recall or re-discovery. The difference may be one of degree, not kind. Once we consider the context of attribution (see Section 2.2), we will be able to give different weights to different resources to distinguish which ones (e.g. a brain, some food, a little time, etc) are usually relatively trivial, and which ones aren't (e.g. a cheat-sheet, years of study, etc).

The potential-parameter furthermore allows us to acknowledge that abilities always take time or resources to produce (subjects don't exist in a vacuum), while also giving us the tools to distinguish the higher quality of ready abilities over strenuously rediscovered ones (all else being equal). So the concept of potential allows us to value people with ready abilities without discrediting the understanding of people who, for instance, need to (re)discover them. This does justice to an observation made by Carter & Gordon (forthcoming):

> "In addition, those with rich objectual understanding will have a certain kind of ability that individuals who ceteris paribus lack rich objectual understanding lack—viz., the ability to easily and accurately piece together new items of information that constitute part of the subject matter. For instance, one with rich objectual understanding will easily see how these new items stand in relations with others within the subject matter"[67] (Carter & Gordon, forthcoming, p. 11)

Before I end, I'd like to draw a similarity between potential and economy. They both focus on the amount of resources required for the appropriate abilities. But economy is focused on the amount of resources used (and allowed to be used, meaning we can focus on the subject as is), and potential is focused on the amount of resources that need to be added (meaning we focus on the subject as it would be). Potential is to economy what robustness was to range.

[67] While they suggest an explanation relying on coherence-making mental relations, I have an ability-based approach to express the same idea.

Furthermore, there is a link between potential and range, as well as robustness. The circumstances of the range parameter may involve resources (e.g. the subject performs appropriately in the range of circumstances that includes a particular resource), and the difference in resources of potential may be expressed as circumstances (e.g. the subject will be able to perform appropriately if we add resource-containing circumstances), so the two parameters are linked, even though their measure is distinct. While range keeps the epistemic subject constant, and considers the amount of salient circumstances under which the appropriate acts present themselves, potential varies the subject and considers the amount of difference in resources which would bring the ability about. In that sense it also has an conceptual similarity with robustness, except that robustness focuses on the subject *retaining* an ability through circumstances, and potential focuses on the subject *gaining* an ability through resources (which could be expressed through circumstances which involve resources, although that's a bit of a detour way to conceptualise it).

On Degrees & Thresholds

Once a subject is attributable with understanding, this does not entail that its epistemic standing is of equal quality as that of any other subject with understanding. Some people's understanding is wider, stronger, or better. It is largely agreed that understanding is not binary, but comes in levels or degrees (e.g. Ylikoski, 2009; Hills, 2015; Baumberger et al, 2016; Van Camp, 2014; Kelp, 2015; Skemp 1976). That being said, it seems counterintuitive to expect a straightforward quantification of understanding on a single numeric scale. Nevertheless, we can readily talk about quality of understanding even in the absence of such quantification techniques, as long as we have a systematic way to distinguish quality or dimensions of quality. And this is something that the parameters can supply. They don't give us a tool to measure the precise quantity of understanding, but they do give us a tool to helps us clearly indicate which things warrant a better understanding and which things can undermine it, and so they can help us decide what to look for and where. Even if we can't express our assessments in absolute numbers, we may find ways to assess one subject's understanding as compared to another.

Even though most authors acknowledge the degrees of understanding, few of them address them as explicitly as was done here. According to Ylikoski (2009), the degrees of understanding are linked to understanding different *aspects*, and to the degree of *control over the phenomenon* (Ylikoski, 2009). Under my account, we may conceptualise these as a subset of scope and sensitivity, respectively. Hills (2015) also links the degrees of understanding to (cognitive) control. These are vindicated by, for instance, the amount of what-if questions that can be answered, which under my account corresponds more closely to the sensitivity parameter. Kelp (2015) measures the degrees of understanding in terms

of the distance from maximal understanding (which I will address in Section 2.3), which is conceived of as being comprehensive and maximally well-connected knowledge. Similarly, Van Camp (2014) conceptualises the degrees of understanding through knowledge of relations and the knowledge structure. Both Kelp and Van Camp mark understanding as knowledge structures, which, depending on how they are conceptualised further, may be open to the objections we raised in Chapter 1, or require further spelling out before they can conceptualise the degrees of understanding. In line with my first chapter, I have conceptualised the degrees of understanding purely in terms of abilities: how many different ones there are (scope), how *sensitive* they are to variations (situational responsiveness) or detail (accuracy), how *stable* they are across or through circumstances (range and robustness) and how *efficiently* they were produced (economy) or could be produced (potential). These parameters dictate the dimensions and degrees in quality of understanding.[68]

We now have at least a rough way to conceptualise what a quality of understanding may look like. But we started this section saying that sometimes the answer to "does S understand?" is quite simply "yes" or "no". How can we account for this, given that the parameters and dimensions are focused on degrees? There is an easy way to incorporate this - namely by using a threshold. Each dimension or parameter may have its own threshold. The notion of thresholds is, of course, by no means a novel idea. In fact, it has already received large agreement in the literature. Furthermore, the point is often made that attributions of understanding vary contextually, be it through the decided threshold (Hills, 2016; Van Camp, 2014, Baumberger et al, 2016) or the actual abilities that would make up the content of these parameters (most notably, de Regt, 2019 allows contextual variation in his intelligibility factors). And this contextual variation of thresholds and content, is where we move to next.

2.2 Context of Evaluations

Which kind of abilities, which circumstances of deployment, which forms of efficiencies and which thresholds are the salient ones for attributing understanding? Up until now, we've more or less assumed that the value or salience of what's inside the parameter can be specified with a single answer that enjoys universal agreement. Unfortunately, and quite unsurprisingly, there is no agreed single universal standard that can clarify all attributions of understanding. And if we try to avoid the varying salience by being all-inclusive, the parameters would cover too much ground to explain the quality of any particular understanding-ascription. But the problem runs deeper, because it is not just that some contexts place a different threshold on how good is good enough, but they also place a

[68] In any given subject and/or topic, we may experience trade-offs between one dimension and another. This would be an interesting line of research, but doesn't concern us here.

different focus on what's most salient. And the problem runs deeper still, because sometimes different contexts contradict each other about what is salient by disagreeing about what is appropriate and what's downright inappropriate. In this section, I will discuss how to conceptualise these contextual variations.

On Context of Attribution and Contextualism

While talking about the dimensions and parameters that determine the quality of understanding, I've sometimes been hinting at the idea that their content and degree may vary with the context of attribution, without explaining what that would entail. It is now time to take a closer look at the role of a context of attribution in determining that content and degree.

To avoid confusion, it would be worthwhile to have a short reminder about what I mean by context, because the word "context" has been used in a variety of ways. The "it depends on X" of contextualism is not the same "it depends on X" as that of subjectivism, the pragmatic reading of epistemology or the modal approach (even though all of them may play a role in determining the context). All of the previous terms have been referred to in some discussions as "contextualism," so I want to be clear on what I do, and do not, mean with contextualism - and what I mean when I use the other terms.

When I talk of subjectivism, I mean that what is true (e.g. whether something feels pleasant) depends on the subject's assessment of the topic in question (e.g. "this feels pleasant for me") - this therefore only applies to claims that relate to a subject's assessment (e.g. how the subject feels about something). Sometimes this position gets overblown into the strawman of complete relativism, where every possible claim is true or false if and only if the subject thinks it so.

When I talk of the pragmatic (see Section 1.1), I mean that what is true (e.g. whether understanding is attributable or an argument is explanatory) depends on the subject we're implicitly or explicitly targeting - this therefore applies only to claims that directly or indirectly involve subjects, such as whether a subject understands or whether an argument is explanatory to a subject. Hempel sometimes conflates this with subjectivism (see Section 1.1).

When I talk of modality and counterfactuality (see Section 1.4), I mean that what is true (e.g. whether a subject has an ability) depends on what happens in actual circumstances, but also what would happen in counterfactual ones (e.g. whether a subject acts under various circumstances and not just

those that have obtained) - this therefore applies only to claims that go beyond the actual circumstances.

But when I talk of *contextualism*, I mean that what is true (e.g. whether a certain subject merits an understanding-attribution) depends on the interests of the attributor, or more broadly, the context of attribution (e.g. whether she satisfies the epistemic abilities that are valued by the epistemic practice in question) - this therefore applies only to claims that have no single universal standard. This is what has been called *attributor contextualism* (Greco, 2008). It entails that:

> "The truth-value of sentences of the form "S knows that p" (and the like) varies with the context of the speaker of the sentence. That is, for the very same S and p, at the very same time, a sentence of the form "S knows that p" can be true relative to one speaker context and false relative to a different speaker context." (Greco, 2008, p. 417)

What Greco says of attributor contextualism for knowledge also holds true of attributor contextualism for understanding. The same subject can be attributed with understanding or denied understanding by different parties, without inconsistency - given contextualism. So where does the contextualism come in? Well, there are a variety of ways in which contextualism plays a role in our attributions. So, let's take a look at some of them.

First of all, it must be noted that we have different standards (not just in thresholds, but in abilities, circumstances or efficiency) for research mathematicians than we do for secondary school students. Khalifa (2013) argues that if the standards of quality for explanations are context-sensitive, then so must the standards of quality for understanding-attribution. Let me remind the reader that the argument here is not that understanding is contextual because explanatory value may vary with the recipient (i.e. is pragmatic), but because the standards of understanding may vary with the context of attribution. As an example, Khalifa contrasts lay people from cutting edge scientists. What's different between them is not just which kinds of explanations are pedagogically effective, but which kinds of explanations are scientifically salient. This is not a pragmatic variation, but a contextual one. What we laud in lay people, we would find inappropriate for experts. Wilkenfeld (2013b) makes the same point:

> "The understanding of someone who got a 4 on her AP calculus test might clearly count as understanding when being evaluated for one job opportunity (e.g. high-school summer intern), clearly not count when being evaluated for another (e.g. professorship in an elite

> math department), and neither clearly count nor not count when being evaluated for yet a third (e.g. admission as a student to an elite math department). The target notion thus exhibits sensitivity to context." (Wilkenfeld, 2013b, p. 130)

However, if this was all there was to a context of attribution, all we needed from contextualism would be a contextual threshold. One for each stage of development. But this may conceive of understanding as too linear. Even if the context of the expert is universally considered in a lot of ways to be the most expedient one (more on that later), it is certainly not the only context that is relevant. This is especially clear to see with cases where learning is not a linear development of piling up further information and abilities (and we'll probably find such cases in most fields). An act or ability that would be considered appropriate at one stage of the learning process, would be considered inappropriate (e.g. inaccurate or wrong) at a higher one. For instance, learning models that are either idealised or downright misleading can nonetheless help students to either appreciate newer, more accurate models or make the new model easier for them to learn.[69] From the context of experts, learning these models would have to be assessed as a way to lose understanding (because all of their answers are now strictly wrong, where before some of them were right). And this would, unfortunately, remain a valid assessment even if this is the same road that the expert took to get where she is now.[70] On pain of excluding such students from understanding-attributions (for any context), we need a way to express what is different - and I propose it is the context of attribution.

A second example of shifts in context can be noted in how we look at the women and men of history. We now have different standards (not just in the threshold, but in the kind of abilities, circumstances or efficiency that are salient) than we did in Hypatia's day (5th century BC) or that of Elizabeth Fulhame (18th century). Consider Elgin's (2007) example of Copernicus (made in the context of factivity rather than contextualism):

> "A central tenet of Copernicus's theory is the contention that the Earth travels around the sun in a circular orbit. Kepler improved on Copernicus by contending that the Earth's orbit is not circular, but elliptical. Having abandoned the commitment to absolute space,

[69] Examples are the number-line metaphor in mathematics, the dichotomy of "healthy" and "unhealthy" in nutrition, atoms as discrete, solid building-blocks in physics, people as rational consumers in economics, humans as descendents from apes in biology, etc

[70] This is not to say that pedagogy always has it right in distinguishing useful shortcuts from retrograde detours, but this requires a discussion about the most appropriate pedagogical contexts of attribution, not which single development-less context of attribution is the universally correct one. (And even within a single stage of development, we may find arguments for several competing contexts of attribution).

> current astronomers can no longer say that the Earth travels around the sun simpliciter, but must talk about how the Earth and the sun move relative to each other. Despite the fact that Copernicus's central claim was strictly false, the theory it belongs to constitutes a major advance in understanding over the Ptolemaic theory it replaced. Kepler's theory is a further advance in understanding, and the current theory is yet a further advance. (...) With each step in the sequence, we understand the motion of the planets better than we did before. But no one claims that science has as yet arrived at the truth about the motion of the planets. Should we say that the use of the term 'understanding' that applies to such cases should be of no interest to epistemology?" (Elgin, 2007, p. 37-38)

Even though the context of attribution has moved on since, that does not make the old context of attribution epistemically sterile in considering these women and men of history.[71] de Regt (2017; de Regt & Dieks, 2005), who also defends a contextual account, has also talked of shifting standards across history. He observes that intelligibility standards (see Section 1.3) have been subject to change or development based on the availability and acceptability of conceptual tools and the preferability of metaphysics (because the scientific community decides which tools and skills are available and therefore required to achieve understanding). On pain of excluding Hypatia, Copernicus or Elizabeth Fulhame (and some of their contemporaries) from understanding attributions, we need a way to express that difference - and I propose it is the context of attribution.

One last example which is worth bringing up is that of variations within contemporary science. Because even within a single field, we can meet with different standards of what is salient (not just in threshold, but in abilities, circumstances or efficiency). For example:

> "In a class on meta-logic meta-theoretical inferences will be relevant, and someone who can make more meta-logical inferences and produce more meta-logical proofs will count as understanding better even if she is not particularly skilled at object-language derivations. In an introductory course on proof construction, the ability to construct object-language proofs will be relevant and a student who can perform more kinds of object-language proofs will count as understanding better, even if she cannot say anything about the meta-theory behind her behavior." (Wilkenfeld, 2013b, p. 130-131)

[71] This is not to say that any stage in history is always part of a step towards better science instead of a retrograde detour, but this requires a discussion about the most appropriate historical contexts of attribution in science, not which single ahistorical context of attribution is the universally correct one. (And even within a single stage in history, we may find arguments for several competing contexts of attribution).

What's more, these different perspectives may actively disagree not just about what is most salient, but about what is salient at all. Constructivist mathematicians reject the validity and use of the reductio ad absurdum proof (of which there are plenty), whereas the formalists would reject any proof that is not completely formalised (thus deeming most mathematical proofs in mathematical practice as incomplete). (Lakatos, 1976) On pain of inconsistency, or excluding all but one set of saliences for understanding attributions, we need a way to express that difference - and I propose it is the context of attribution.

At this point, it becomes important to note that contextualism does not mean everyone gets to make up their own scientific methodology. The debate about scientific methodology and scientific aims is either a debate about the suitable context(s) of attribution, or a debate that informs it. As such, the proposed answers in that debate will put constraints on (and give direction to) which context is appropriate to judge people's understanding from. So the problem with astrology (a staple example of pseudoscience) is, as I mentioned earlier, that it either devalues clarity, predictive power, falsification, etc (staples of scientific methodologies), and thereby faces criticism that its context of attribution is not suitably scientific, or it does value such things, and its practice will fail even by their own context of attribution. Allowing different contexts of attribution is no more harmful to science than it is to allow sociologists to use different methods than physicists in how they do their research and, by extension, on what basis they evaluate their members' epistemic standing. What allowing different contexts of attribution does do, is allow us to chart the ways in which they differ more explicitly.[72]

As such, our approach to understanding could do with a contextual spotlight that specifies, within each parameter, that which is salient to it. The way to deal with this is by allowing the context of attribution to give more or less weight to specific kinds of abilities, circumstances or efficiencies, along with the option for thresholds. If the context of attribution is supposed to address variance, we should be able to handle the variety of appropriate understanding-ascriptions in a principled manner. That's what the rest of this section will expand on. Each parameter will be given the same two facets of contextuality (see also Wilkenfeld, 2013b[73]): (i) a weight of relevance given to each parameter-component (what the contextual light shines on) and (ii) a (possible) threshold, which determines the degree to which those parameter-components must be present.

[72] For a paper that addresses the worries of disappeared objectivity or hollowed out topics, see (Greco, 2008).

[73] "There are at least two dimensions of variability—which sorts of attributes matter to determining whether something counts as understanding will vary contextually, as will the degree to which those attributes must be present to be above some threshold of understanding." (Wilkenfeld, 2013b, p. 130)

To be clear, I am not claiming that we can always chart the context of attribution through a complete list of aims and uses which a subject is intended to display. This is not a way to exhaustively chart a scientific practice or paradigm. What I'm claiming is that any validation, or discredit, of understanding-attributions (and disagreements thereof) should translate into a contextual interest that the subject meets, or fails to meet, within a context of attribution (but perhaps not within another). I am claiming that this is just a fruitful way of making those claims more explicit (as will be showcased in the relevant examples of Chapter 3).

Scope Interests

The first dimension to consider is that of scope. Scope was the amount of variety in abilities salient to understanding an object X (which can be expressed as its meaning - see Section 1.4). But both the content and threshold of scope (namely, which abilities are salient to understanding object X and how many suffice) needn't be agreed upon by all parties interested in attributing understanding of object X. The same object can belong to different *domains*. As will become clear, I use the word in its most neutral form. The concept of "domain" may therefore be interpreted as broad as "field", or as narrow as "research agenda." But because each domain has their own interests, they also have their own requirements for which uses (see Section 1.4) falls within their scope for understanding its objects:

> *Scope (domain) weights*: which uses connected to X are deemed appropriate, and to what extent.

This contextual light groups salient uses or abilities into one (more or less distinct and explicit) context of attribution. The way these uses or abilities are grouped will depend on several (overlapping) patterns that are valued by a particular practice or sub-practice, a domain. An object X (and some of its uses) may be shared by several domains, but its precise meaning (and therefore its salient uses) will be relative to the domain in question. Each domain could be distinguished or grouped by its field of problems, intended aims, empirical standards, conceptual tools, methodological constraints, logical requirements, standardised symbols, background metaphysics, field of focus, or pedagogical stage. These will dictate which abilities are salient to that domain and for its attributions of understanding.

Even within single disciplines, we can distinguish several domains.[74] The distinction can be due to many different things. The distinction in domains can, for instance, be due to differences in approach.

[74] Instead of carving up different contexts of attribution, one could insist that we duly specify which types of interests are supposed to be served by further specifying the exact type of object through which domain it belongs to. For some circumstances, that is definitely possible ("prove this algebraically"), but it seems more true to practice to gauge the

For example: what it takes to understand a proof of a theorem in algebra involves a different set of abilities than it does to understand a proof of that same theorem in topology or geometry. Domains can also be distinguished by certain decisions of methodological constraints. For example: constructivist mathematicians reject the reductio ad absurdum as a permissible move in mathematics, whereas formalists reject informal proofs. The difference is not in subfields, but in methodological constraints on certain moves in the science game. Differences in domains can also be distinguished by which conceptual tools are deemed valuable. Visualisable theories are often seen as more tractable than abstract ones, but some scientists actually prefer abstract ones over visualisable ones. For instance, most physicists at the time of Schrodinger and Heisenberg preferred Schrodinger's wave mechanics because of their visualisability, whereas others rejected Schrodinger's visualisability in favour of Heisenberg's more mathematically intricate matrix mechanics. (See de Regt, 2019, c7)

Metaphysics can also play a role, as has been detailed by De Regt (2017, c5). This is what was at the core of the dispute between Newton and Huygens (a dispute that has been oscillating beyond their lives). Action-at-a-distance did not fit into the metaphysics that was reigning at the time, therefore the concept of action-at-a-distance was deemed unacceptable. At the time of Newton, Descartes's mechanics was the most successful physical theory and it was based on the specifically mechanistic conception of causality, namely corpuscularism, where contact action was essential. This metaphysical worldview had moreover contributed to scientific progress, so was not contingently accepted. So when Newton relied on action-at-a-distance, this was initially considered as unacceptable and rejected as unintelligible until familiarity, and the success that came with it, vindicated it.

The contrast between domains is even more striking if we focus on the differences between whole disciplines. If you ask why a particular car-crash happened, you will get a different appropriate response from a lawyer, a physicist, a town-planner, a politician or a psychologist. Furthermore, the same phenomenon will quite often not even get the same questions. For instance: an economist is more interested in the question of how the gender wage gap affects the GDP, whereas a sociologist is more interested in the question of why it is women who are paid less and not men. They have a different focus. For either related or additional reasons, they also deal with different epistemological

appropriate context of attribution than to carve up different objects of understanding for every possible variation in interests. Wilkenfeld (2013a) makes a similar point: "We could try to locate the context sensitivity entirely in the specification of an object of understanding, but (...) [u]nless one builds a particular perspective into the specification of an object, the same object— described to arbitrarily fine grain—can be viewed from different perspectives. Hence it will be possible for the very same object to be understood in some contexts but not in others." (p. 1008)

problems. Failing to explain why a company went bankrupt is an open problem for economists, but not so much for sociologists. This may or may not have something to do with their intended aims: is it to explain, predict or control the object of study? And if they predict, what sort of empirical standards are salient to the domain? It is no surprise that the empirical standards of prediction are different for psychologists than they are for physicists, not just in accuracy, but also in the types of evidence that are considered salient.

So by the word "domain," I here mean any of the previously mentioned perspectives that can be used to distinguish the type of act that is appropriate from those that are not (be it because they are inappropriate or merely irrelevant). These perspectives can come from radically different directions, based on disciplines (e.g. psychology, economics, mathematics), subfields (e.g. algebraic, geometrical or topological approaches), target of focus (e.g. atoms, individuals, societies), levels of difficulty (e.g. primary school, secondary school or research-level mathematics), intended general aims (e.g. prediction, explanation, control)[75], preferences of conceptual tools (e.g. a preference for visualisability, mathematical abstraction, causality), field of problems (why did that company fail, how do we keep a company from failing), available metaphysics (e.g. action-at-a-distance, corpuscularism), empirical standards (e.g. computerised data over human intuition, or vice versa), logical requirements (e.g. in mathematics, inconsistencies are more damaging than they are in psychology) or some other way of grouping the appropriate abilities (e.g. the abilities salient to why-questions/how-questions, using algorithms, translating scientific texts or teaching the subject matter to someone else) or combination thereof.

I'm not claiming that the context of attribution regarding scope can always be specified with a finite list of salient uses or abilities, but I am claiming that disagreement over understanding attributions may result from the different weights a context of attribution places on which acts are salient to warrant attributions.

Sensitivity Interests

Next, let's look at the different contextual lights which can shine on the sensitivity parameters. As a reminder, the sensitivity parameters were situational responsiveness (i.e. amount of appropriate

[75] There can be specific non-epistemic aims too, such as receiving more grants, producing more coffee, social equality between all genders. Since all sciences are natural human practices, I am not sure whether there is always a clear-cut distinction between epistemic and non-epistemic aims that define that practice, and both may be salient to understanding in certain contexts.

changes in performance to changes in the object-situation) and accuracy (i.e. degree of precision in performance).[76]

In honour of the famed "what if things had been different questions" that characterised situational responsiveness, I call its contextual light "What if" weights:

> *Situational (What if) weights:* which variations in object-situations, along with their appropriate reactions are relevant, and to what extent.

For example, when it comes to understanding space-trajectories, the what-if of an additional planet might be more salient to NASA, whereas the what-if of an additional cosmological constant might be more salient to theoretical physicists. If one is interested in suitable responses to what-if-things-had-been-different questions (or other variations in object-situations not necessarily phrased as questions), then we must be able to determine which of those questions (or situations) as well as their answers (or responses) are salient, and that's what this contextual light allows us to do. It tells us which variations are more or less salient, depending on the interests of a (sub)practice. Like scope, which acts (and how many) can be considered salient to a single ability is relative to the (sub)practice making the attribution (i.e. the context of attribution).

Secondly, we can have varying interests in accuracy:

> *Accuracy weights*: which types of accuracy are appropriate (when there are degrees to success) and to what extent.

For example: To calculate the square footage of a house, accuracy to a single decimal place may be sufficient, but to produce a flight plan in outer space, NASA will need more than the single decimal place. What degree of accuracy is salient for understanding the position of the sun is different for a navigator than it is for a physicist. [77] But the salience of accuracy may also differ in type. Weather

[76] Situational responsiveness was about the variety of appropriate acts that can make up one ability. As there was no necessary dividing line between one type of ability and one type of act, there was also no such line dividing scope and situational responsiveness. This means that one could express the contextual light of scope through situational responsiveness and vice versa. But it is fruitful to distinguish the two to focus on variations within a single ability.

[77] Scope and sensitivity interests have consequences for the notion of factivity in understanding. For idealisations to be worrisome (or even to count as idealisations), the difference between the acts of what is considered a strictly false and strictly "true" belief needs to be a salient one. Therefore, what distinguishes an "idealized" theory (or belief) which is considered not strictly true, from a non-idealised one, must result in a difference in acts or abilities - for how else would we discern the difference? But the salience of the difference may vary with the context of attribution. For a particular context of attribution, what is inappropriate may simply fall outside of what is considered salient, or be

forecasters may care more about getting short-termed predictions precise than long-term predictions adequate, whereas climatologists (or insurance-agencies) care more about the reverse. Furthermore, acts can not only be appropriate or inappropriate, but can also be somewhere in between - they can be helpful or approximate. For some contexts, being helpful or approximate (instead of entirely successful) is salient, whereas for others only downright success will do.[78]

Stability Interests

Next in line is the stability dimension. As a reminder, the components of the stability parameters were range (i.e. degree of presence in (counter)factual circumstances) and robustness (i.e. degree of presence after (counter)factual circumstances). However, it doesn't just matter that the subject would act appropriately in, or after, any circumstances, but also *which* circumstances that subject would act appropriately in or after. Not all are equally salient. Stability interests are a way to express that salience by giving them extra weight in our assessment (for that context of attribution).

The first contextualist spotlight within stability will target the appropriate range of deployment.

> *Range (deployment) weights* = which types of (counter)factual circumstances, where the same (barring non-epistemic changes) subject acts appropriately, are salient and to what extent.

There's a time and a place for everything, and here we can specify what they are. We not only like our mathematicians to produce proofs, but to produce the proofs that are relevant to the circumstances. Two mathematicians may be equally capable of producing existing and new mathematical results (i.e. have the same *degree* of range), but what if one of them does so when the circumstances call for it (for example: when a colleague or research fund asks for it) and the other does it at random moments (maybe, for example, she frequently decorates her house with proofs)? Although we may attribute them with an equal amount of abilities, it is the former understanding which we value most in the

equally appropriate (e.g. to a meteorologist, it makes no difference whether one uses an idealised or detailed theory of the sun's trajectory). For another context still, the salience may vary in degrees.

[78] It may be interesting to note that weights can be reconsidered with trade-off successes or failures in other parameters. For instance: "classical electrodynamics is to be preferred over a quantum treatment, even though the latter approach would (theoretically) lead to more accurate results. The reason is that a quantum treatment of classical phenomena would be enormously complex while the gain in accuracy will in practice be negligible." (de Regt, 2017, p. 38) Here, the weight of certain degrees of accuracy are lower than the weights in efficiency. In other contexts, this is not the case. In (de Regt & Gijsbers, 2017) it is shown why, in most contexts, (the abilities of) Newtonian gravitational theory carries more weight than (those of) Einstein's general relativity. So the context of attribution may shift its salience in one aspect to help that of another.

context of research-mathematics, because they are performed under the valuable circumstances. Deployment weights are a way to give more weight to those valuable circumstances, where needed.

The salient circumstances will tend to be those that are deemed fair to the agent (e.g. room temperature is generally a more relevant temperature than anything significantly higher or lower), even allowing peculiar circumstances, should they be beneficial but otherwise harmless (e.g. working without distractions, coloured overlays, etc), and/or circumstances which are most beneficial to the practice (e.g. when asked or paid with monthly salary), but excluding circumstances which would be unexpected (e.g. performing while in a hurricane is less of a concern for mathematicians to give it much salience) or damaging for the agent (e.g. extreme heat or inside a vat of nuclear waste).

The range parameter (or even stability dimension) doesn't often get mentioned, unless very obliquely. There are exceptions. For example: Sierpinska says understanding is the potential for acts-of-understanding in a context *where it is necessary*. (Sierpinska, 2005) More notable exceptions are Goldin (1998) and Wilkenfeld (2013a; 2013b):

> "There is at least one other dimension to our common-sense notion of understanding: the extent to which one understands varies not just with what one can do, but the conditions under which one can do it." (Wilkenfeld, 2013b, p. 1008)

He continues with an example:

> "Suppose Vir is a brilliant psychologist but markedly socially inept, whereas Londo is a social butterfly who could not put two sentences together to explain how people behave (much less why). When seeking a speaker for a conference on the psychology of party-goers, one would clearly want Vir; in a context where we are looking at applications to speak from Vir and Londo, it would be fair to judge that Vir understands people, whereas Londo does not. However, matters differ sharply when sending out invitations to a party. We are well aware that Vir will be awkward and never know how to interpret people or how to respond appropriately. In that context, it seems reasonable to judge that Vir does not understand people, whereas Londo does. Differing contexts thus exhibit discontinuous criteria regarding what counts as understanding, even of the same object." (Wilkenfeld, 2013a, p. 1008)

Because Wilkenfeld's account isn't ability-based, his example is one where the understanding seems to be about the same thing, but the type of abilities (and therefore their scope) are different. But this is not crucial to the example. Londo and Vir could have the same type of abilities, but merely display them in different circumstances, and the example would be equally pressing. Whether they understand people is a question which can only be answered if we explicitly or implicitly know what sort of circumstances they need to display these abilities in. Both the context of attribution where one favours Londo and the context where one favours Vir (as well as a broader one which incorporates both) are valid, but they suit different interests. Later, Wilkenfeld continues more to the point (in my view):

> "Of course in the counterfactual scenario where Vir is sitting in a comfortable chair at a conference discussion panel, he would be able to represent the theoretical party-goers very well and make all sorts of interesting inferences regarding their behavior; however, whatever salience amounts to, that counterfactual is not typically salient during an average party. We could imagine cases where it becomes salient, but those are the very same cases where the correctness of judgments regarding his understanding seem to shift—were someone to ask specifically about Vir's academic credentials, one could very well imagine being told, even as Vir commits yet another faux pas, that Vir really understands people." (Wilkenfeld, 2013a, p. 1009)

This example also shows that what *changes* the salience is due to the context of attribution, not the circumstances where Vir finds himself in (except to the extent that the circumstances dictate a different context of attribution). (Wilkenfeld, 2013a) Circumstances are sometimes also referred to as context, but this sense of "context" is primarily intended to convey the salient "object situation", not the salient environment for deployment.

We needn't expect a range to be specified exactly or exhaustively. Even a vague description of range would do. One's competence, as Goldin (1998) puts it, is related to one's ability to "perform a task some of the time, under conditions which are partially but incompletely specified" (p. 147) And that specification, even if partial and incomplete, is what deployment weights are concerned with.

A possible objection to the contextual approach (as presented) could be a case where someone only acts appropriately outside of the salient circumstances. Imagine we are trying to gauge whether Beth, who is an employed meteorologist, understands anything about meteorology at all. For whatever

personal reasons, she has decided to act extremely contrarian. She does not respond appropriately under any of the circumstances that we consider to be salient. When a reporter asks her to explain some of the causes of the climate crisis, she scowls, when she gets the assignment to forecast tomorrow's weather, she writes "I don't know," and whenever we observe her handling any equipment, she breaks them. We consider these circumstances to be salient to gauge her understanding, but under those circumstances she fails miserably. However, in her private life, Beth brings an umbrella for the upcoming rain, she repairs one of her anemometers and is trying to persuade her husband that going vegan and avoiding fossil fuels will help save the environment. What this shows is not that the earlier context is wrong, but that there are multiple possible contexts of attribution. From the vantage point of a context which puts salience on abilities in a work-environment, she lacks the relevant abilities.[79] But from the vantage point of a context which puts salience on abilities in everyday circumstances, she clearly does display some of the relevant abilities. So the counterargument to this counterexample is that the conclusion should not be "lose the context of attribution," but "reconsider your context".

Next, our contextual light can also target the salient circumstances a subject needs to be able to face up to. Here, we are considering what is salient within the robustness parameter:

> *Robustness (rationality) weight*s = Which types of circumstances, after which the subject needs to continue to act appropriately, are salient and to what extent.

Rationality weights are a way to decide which circumstances or events the subject must definitely be able to face. There are some obvious candidates for rationality weights, namely that it is usually salient that the subject not to be too forgetful (after a little time or distraction), or too easily swayed (by bad evidence or testimony). Here the contextual light can specify not only which robustness is especially valuable (e.g. not changing one's mind too easily), but also exclude which especially *isn't* valuable. An example of this is being convinced to change one's mind (and thus performance) due to good evidence or some other form of justification. What counts as good evidence or justification will depend on the epistemic standards of a practice (at a particular time) and *rationality interests* are intended to cover those standards.

[79] This example is misleading for a second reason. It must be noted that, because she has made the decision to act contrarian, she will have a reason for this. (e.g. because she is severely underpaid), which entails that an analysis of her counterfactual range would reveal that in a counterfactual world where this reason is absent (e.g. a world where she is paid appropriately), she does display the relevant abilities. I'll come back to these types of arguments in Chapter 3.

Perhaps a potential weakness of the robustness parameter is that someone who is soon to be afflicted with Alzheimer's or who is terminally ill would invariably score lower in robustness, and therefore in understanding, than someone who is not. This may be appropriate for comparing candidates in job-interviews, where long term robustness is particularly salient, but in most other contexts of attribution, how close they are to death or Alzheimer's is not as salient in comparing epistemic subjects as is their short-term range and robustness. The context of attribution can quarantine this oddity by denying the circumstances of death or Alzheimer's with salience.

System Efficiency Interests

Now we've come to the last dimension, namely that of system efficiency. When we shine our contextual light on what is salient in system efficiency, we'll do so within the economy parameter (where the appropriate act uses a minimum of saliently allowable resources) and the potential parameter (where the appropriate act obtains with the addition of a minimum of resources or events). System efficiency interests are a way of giving weight to the salience of particular resources or events which are salient in displaying the appropriate acts.

First we need a way to express the desired salience within the economy parameter. Without a contextual focus picking out the desired salience, the threshold of economy is especially difficult to characterise. What should we be most economical about? Do we express the concept economy in terms of size? Clearly not, because someone using a small, sophisticated and (very expensive) calculator would then be more economical than someone using a large standard-issue calculator. It is not just the sum amount of time, space, energy,... that matters, but how these resources are assessed by the practice doing the attributing. Here's a rough characterisation:

> *Economy (resource) weights* = the particular resources (incl. events) one is willing to allow to be used to consider the abilities achieved, and to what extent.

Wilkenfeld (2013a) doesn't get into context-sensitivity of efficiency, even though he does hint at economy with regards to the shy Vir:

> "When Vir is surrounded by loud and bustling party-goers, he is simply not capable of updating his representations well enough *or fast enough* to be able to interact effectively with the world." (p. 1009, italics added)

The implication here (presented in the language of representations) is that Vir might be able to respond appropriately, but that his responses would be too slow to be salient. This is a form of inefficiency, where the resource is time, and Vir is simply taking way too much of it. Other examples of types of resources mentioned earlier were a computer, a calculator, a sheet of rules, a library of information or food and drink, etc. Resources also include whatever makes up individuals, such as brains, but since most subjects always carry their full brain with them everywhere they go, we don't usually divide economy up into specific brain-resources (e.g. how much of the brain is used or how many lobes were employed). If this should ever change, however, then perhaps this content of the economy-parameter would become more salient. Until then, this parameter will be largely concerned with the circumstances surrounding those brains. Overall, I expect that the contextual light for economy will shine most brightly on the use of resources that are easy or cheap to acquire or produce, and shine less brightly on those that are not. In most contexts, this entails that the subject uses her brain and her body, a little time and any form and amount of sustenance and drink to survive. In the context of, for instance, a math exam, these are exactly the sort of things that are usually allowed. Most resources that go beyond it (e.g. textbooks, cheat-sheets, collaboration with others[80]) are not permissible even if they were to result in the appropriate abilities. In research mathematics, I imagine the use of an existing body of research can be expected and is therefore permissible as mere background conditions. The contextual concept of resource weights is simply a way for making that explicit. And, as always, we can specify the desired degree of appropriate economy with a threshold.

Being able to shine a contextual light on resources does also open the door for a practice to exclude anything (e.g. a computer program) or anyone (e.g. people of a different gender or a different nationality) they don't like/respect by considering them an unvalued resource. While this is allowable in theory, contextual lights are not an invitation to bigotism or chauvinism. To avoid it, discussions are required about the appropriate context of attribution. As has been argued extensively in the literature, such as that of feminist epistemology, context of attributions that are intentionally or unintentionally bigoted or chauvinistic are subject to the strong justifiable critique of being morally unfair and epistemologically undesirable. But this critique is directed at the inappropriateness of a particular context of attribution, not at the concept of contextualism.

Lastly, we can contextually specify the desired kind of potential. Here is the potential weights concept, roughly characterised:

[80] This is true even if the appropriate abilities of cooperation cannot be reduced to the abilities of the students separate. For an extensive conceptualisation of this possibility, see chapter 5, which is specifically devoted to the notion of collective understanding.

Potential weights = the particular resources and circumstances one is willing to consider in order to assess the abilities achieved with it, and to what extent.

Examples of resources are those mentioned above (such as a calculator), but also events such as having questions answered, studying, tutoring lessons, empirical research, or brain surgery. Here we assume that the resources that weren't used could be used. Or that the resources and events that aren't "part of the subject (yet)" are nonetheless salient to consider. They are often the things one is willing to throw at the subject to unlock the abilities. In principle anyone and anything has some potential if you're willing to go far enough to unlock it. However, sometimes what you have to throw at the subject is a lot more difficult, expensive and complex than it seems to be worth it. Extensive and specific brain-surgery[81], for instance, might make someone understand, but in terms of salient potential it'll be rather much lower than an education, and much lower still than a simple piece of information. As always, a threshold can be added which would specify the degree of desired potential.

On Contextual Determinants

When we are assessing a subject's quality of understanding, it is my contention that we are assessing the degrees of the aforementioned parameters, where the content and thresholds are determined by what the context of attribution considers salient. Thus far, I have motivated appropriateness only by allowing an appeal to the authority of a particular practice. The only normative import of my account has been that any claim, on pain of vacuousness, needs to boil down to appropriate uses, acts or abilities (be it either directly or indirectly) and that these uses, acts or abilities need to flow directly and consistently from the values and aims of the practice in question.[82] But an exhaustive investigation of understanding would preferably go beyond the mostly descriptive "appeal to practice" and be able to say something about how and why a practice would or should deem certain uses, acts or abilities as appropriate. In short, it would discuss *contextual determinants*. To do so is to open the conversation of which context(s) of attribution (which decides the appropriateness of acts, abilities, stability and efficiency) are themselves appropriate. For *scientific* understanding, this can be seen as the problem of deciding whether the methodologies, aims and values of a practice (a context of attribution) are appropriately scientific. But because this depends on the type of practice and how it evolves over time, I believe this is a discussion for which I have neither the means nor the authority to shed much light

[81] It's currently well beyond our means, but if the difference between not understanding and understanding can be found in the way a part of the brain is wired, then it would in principle (with a very large emphasis on "in principle") be possible to directly wire the brain in the appropriate manner.

[82] For example, if a practice values certain kinds of prediction very highly, it needs to value the corresponding acts of predictions.

on. The search for a universal scientific demarcation principle has famously failed, and even within particular disciplines, what makes for good science evolves or branches out all the time.

> "We have seen time and again that the aims of science vary, and quite appropriately so, from one epoch to another, from one scientific field to another, and sometimes among researchers in the same field." (Laudan, 1984, p. 138 quoted in de Regt, 2017, p. 89)

This is not to imply that what is appropriate to a (scientific) practice is (entirely) contingent (nor that it isn't), but what makes an appropriately good science or practice is a question without an easy, universal or timeless answer. And I'm more than willing to leave the discussion of what constitutes good science out of my own hands.

This does not, however, render my account meaningless. If anything, it is a virtue of the account that it can conceptualise understanding, and its evaluation, through the lens of different contexts of attribution while at the same time shifting the problem of what is (or are) good context(s) of attribution to a different discussion.[83] Furthermore, it helps that discussion by making a context of attribution more explicit, and thus open for critique (e.g. of inconsistencies and/or selective blindness[84]). But more can be said about the appropriateness of contexts of attribution, and I will do so briefly here. I will focus on (a) scientific demarcation and (b) fairness as contextual determinants.

(a) Contexts and Scientific Demarcation

Following de Regt (2017), who focuses more on the successful constructing of models to understand than on assessments of understanding, I will briefly touch on some contextual determinants to show that allowing varying contextual lights is part of good science and being open to contextual lights doesn't invite a free-for-all. Even in the absence of scientific demarcation-criteria, we can offer some values that influence whatever does demarcate a theory as scientific. The two main ones are *internal*

[83] To legitimise science in spite of variations, de Regt (2017, c4) appeals instead to a distinction between the macro (whole), meso (community) and micro (individual) level of science. The different practice of sciences across disciplines or history share the macro-level aims of producing knowledge that is supported by experience, but differ in their meso level in how to do that and what about. I'm inclined to agree, but it seems to me that the meso-level discussion of what makes for a good particular practice of science is more important in its legitimation as a science than its macro-level similarity with other sciences.

[84] In fact, in Chapters 4 to 6 I will argue in favour of non-human understanding entirely based on why certain contexts of attribution are biasedly inconsistent or unduly blind. If one wishes to insist that the targets of understanding attributions can or should only be human individuals (and not extended systems, groups or artificial systems), then one either fails to consider the abilities of individuals and non-individuals with the same eyes and one is biasedly inconsistent, or one refuses to let any eyes fall on the macro-systematicities of non-individuals and one is therefore unduly blind.

consistency and *empirical adequacy*. Neither of them, however, are straightforward or without trade-offs. For instance:

> "Empirical adequacy is far from straightforwardly and unambiguously applicable as a criterion: the usual situation in scientific practice is that theories or models fit the empirical evidence only partially and that a choice between different theories or models involves value judgments about which evidence is considered most important." (de Regt, 2017, p. 38)

Furthermore, both internal consistency and empirical adequacy can trade or interact with *additional values*, be they general qualities (e.g. simplicity, consistency with background knowledge) or specific conceptual tools (visualisability, causality, unification). de Regt proposes that the choice of values has something to do with contextual metaphysical preferences and with intelligibility. *Intelligibility* is the aggregate value of all qualities that are relevant to facilitate the use of a scientific theory in a scientific practice. The intelligibility of the theory is best evidenced by the subject's ability to recognise qualitatively characteristic consequences of the theory without performing exact calculations. Therefore, intelligible theories are valuable, not because they are subjectively more congenial, but because they are pragmatically more fruitful. But what is fruitful for one group of scientists isn't for another. The theories that are most intelligible, are the ones of which the theoretical qualities are most attuned to the scientist's skills to use them, so what is deemed intelligible can vary through time or across disciplines. Different disciplines, communities, historical periods offer different background knowledge, and characteristics of entrenched theories, so they hone different skill-sets. Conceptual tools, such as visualisation, causality and mathematical abstraction, can render scientific theories more or less intelligible, depending on the preferences, skills and background knowledge of the scientist or the epistemic framework of the discipline. Visualisable theories are often seen as more tractable than abstract ones, but some scientists actually prefer abstract ones over visualisable ones. This was at the heart of the dispute between Heisenberg's matrix mechanics and Schrodinger's wave mechanics. To most physicists, Schrodinger's wave mechanics was deemed more intelligible (and thus more fruitful and therefore scientifically valuable) thanks to its visualisability, whereas adherents of Heisenberg's more mathematically intricate matrix mechanics rejected the need for visualisability. (See de Regt, 2019, c7) In short, even if the fruits are objective, what is the most scientifically fruitful varies contextually with what's most pragmatically within a subject's reach to pluck.

> "The fact that there is not one generally accepted interpretation of quantum mechanics and that physicists belonging to different schools claim to understand the theory by means of their own conceptual framework (and often accuse other approaches of unintelligibility), illustrates once again the contextuality of scientific understanding and the fact that understanding is not bound up with fixed explanatory categories." (de Regt & Dieks, 2005, p. 162)

Secondly, de Regt (2017, c5) also shows how intelligibility can play a role in metaphysics by not only determining which conceptual tools are available, but also which are *acceptable* and, more crucially, which are not. This was at the heart of the dispute between Newton and Huygens. For Huygens, action-at-a-distance was unintelligible, having learned to understand the world by mechanistic principles and models, and the notion of action at a distance flatly contradicted the principle of contact action. But as scientists acquired the skills to work with Newton's theory, the theory could be vindicated by its subsequent successes. (de Regt, 2017, c7)

Lastly, I'd like to add that different contexts of attributions can serve different epistemic interests.[85] I don't think we need a justification for why lawyers may focus on different aspects of a situation than psychologists or physicists, or why meteorologists favour efficient predictions (e.g. predicting when it will rain) over control (e.g. making sure one doesn't get wet), while the study of medicine cares more about control (e.g. keeping someone alive) than efficient predictions (e.g. precisely predicting how long someone will stay alive for). Van Fraassen (1980) believes pragmatic aims such as these only contextually vary "human concerns", not epistemic ones. De Regt (2017, c3) disagrees because "science is a human activity and it makes no sense to speak of "the aims of science" in the absence of human concerns" (p. 125). I agree with de Regt. But whether you agree with de Regt or van Fraassen, the point still stands that there is a contextual variation of what is salient. And whether something is salient depends on whether it serves the contextual interests, be they epistemic or human.

Furthermore, even if we can distinguish epistemic aims (e.g. prediction) from non-epistemic aims (e.g. getting more publications and a higher h-index), there can be trade-offs between multiple epistemic aims which one context values more than another (e.g. prediction over control or vice versa), with no single context that can (or should) incorporate both. What's more, there's no reason to assume that all epistemic aims will be satisfied with the one true context of attribution. Sometimes, pluralism is a

[85] See also (Delarivière, Frans & Van Kerkhove, 2017).

virtue, not a vice. (See, for instance, Chang, 2012) To borrow a phrase from Elgin (2004): "Context provides the framework. Purposes fix the ends." (p. 121)

In short, the most fruitful scientific theories are those that stroke most successfully with the accepted metaphysics and the available conceptual tools, which fit best with intuitions, rely on existing skills and best suit the epistemic aims. The context both shapes and is shaped by them, though not contingently so. And unless the sciences settle on a single, timeless and universal monist view of how to approach, conceptualise and focus on all of its objects of study as well as which single unambiguous set of aims they wish to satisfy with that endeavour, we need different contexts of attribution for the plurality within the sciences - even if it requires a justification for each of the contexts of attribution as being appropriately scientific.

(b) Contexts and Fair Expectations

Focusing on contextual interest also brings attention to the fact that most of our contexts of attributions are based on neurotypical subjects. Making this explicit makes it easier to see why it could be advantageous to vary the context of attribution for neurodivergent subjects, such as people with dyslexia, autism, ADHD, etc. The circumstances under which people with dyslexia perform well (e.g. colour overlays) may be different from those of others. And yet it is likely that what is considered a salient range for assessing a subject is, at least partly, based on what is fair for a neurotypical subject. A similar point can be made about scope and sensitivity (as well as system efficiency). It is likely that the abilities that come more naturally to neurotypical subjects are weighed higher (because they are more readily expected) than those that come more naturally for neurodivergent subjects.

> "Consider a dyslexic student. Universities (at least in the United States) make special allowances for students with such conditions as dyslexia precisely because it is recognized that they cannot always perform the tasks we come to expect of people who understand the material. One might be cautiously optimistic that there are other tasks one could ask them to perform such that their performance would be reflective of genuine understanding (...)" (Wilkenfeld, 2013a, p. 1012)

Kelp (2015) also draws attention to this, but claims (to some degree) that this poses a problem for ability-based accounts:

> "[D]yslexic agents may understand a certain phenomenon even if they are unable to perform any task whatsoever that would be reflective of it. They will do so whenever they know enough about it to be such that they would (be sufficiently likely to) successfully perform the contextually determined set of tasks if they were to have the skills needed to do so and to exercise them in suitably favourable conditions" (Kelp, 2015, p. 22)

From the vantage point of my counterfactual and contextual account, it is clear why "would (be sufficiently likely to) successfully perform" is actually just another claim about (counterfactual) abilities. The relevant question then is which circumstances (or "conditions") and how many are contextually salient for a fair context of attribution. What marks the understanding in the neurotypical and neurodivergent subject are both abilities, but which range is considered fair for a subject with dyslexia is different from a neurotypical subject.[86] A discussion about the context of attribution draws attention to this and raises the question of whether this is actually fair or desirable. Rather than assume that the context of what is to be considered salient should be based on what is fair for the typical subject, we need to justify the context of attribution as appropriately fair, given the variation of human strengths, limitations and needs. In short, having to justify the context of attribution makes it clearer why we do not automatically need to assume that what is neurotypical should be neuronormative.

2.3 Evaluations of Quality

We have considered the dimensions and parameters that determine the quality of understanding, as well as how a context of attribution varies their content, but what we have not yet considered or helped to conceptualise is some of the practical concerns in evaluating these dimensions and parameters. In this section, I will conceptualise the evaluation side of understanding by addressing some of its aspects. I will label a common misevaluation, introduce a distinction between direct and indirect evidence, point to some limits of characterising a context of attribution, show how contexts of attribution can also handle kinds of understanding, and consider what it might mean to have complete understanding.

Evaluation of Competence & Misevaluation of Kludges

As I have said before, it is my contention that when we are assessing a subject's quality of understanding, we are assessing the degrees of the aforementioned parameters, where the content

[86] Kelp may mean the same thing in including "suitably favourable conditions" and "sufficiently likely," except that he does not expand on these contextual variations of salience with the same care as he does contextual thresholds.

and thresholds are determined by what the context of attribution considers salient. Unfortunately, however, we never have the benefit of witnessing the full scope, sensitivity, economy, potential and (counter)factual range or robustness of that subject, so we always have to evaluate the quality of understanding based on a limited set of actual performances. I contend that most understanding attributions are estimations, based on limited evidence, of how the subject would fare beyond what has been discerned. They are assumed generalisations. For instance: we assume that a subject can repeat her appropriate acts if she's done so before, we assume the subject has a wide range of circumstances if she performs appropriately under the several different circumstances tested, we assume a subject has a wide scope if the different abilities tested were all present, we assume she is sensitive to object variations if she has adequately varied her acts to the situation, we assume she is efficient if she has always been quick and successful without help. Any of these evaluations are based on limited evidence, and assume (until proven otherwise) that their applicability stretches beyond those precise circumstances, abilities and efficiency tested. This means that our estimations of a subject's understanding are open to over- as well as under-estimation.

To make clearer some dangers of over-estimation, I will borrow the term "kludge" from the computer (and engineering) sciences. A *kludge* is jargon for a cheap trick, a quick-and-dirty solution or short-cut tactic to have something perform as desired. (Sharples et al, 1994) The reason it is cheap, quick-and-dirty or a short-cut is because the success of its performance is strictly limited. But this entails that an assessment of understanding that only evaluates a limited set of evidence is always open to incorrect generalisations. Kludges can make computers appear cleverer or more complex than they actually are. But kludges can also make humans appear to have more understanding than they actually do. Skemp (1976) gives a nice example of one (although he doesn't call it a kludge):

> "Recently I was trying to help a boy who had learnt to multiply two decimal fractions together by dropping the decimal point, multiplying as for whole numbers, and re-inserting the decimal point to give the same total number of digits after the decimal point as there were before. This is a handy method if you know why it works. Through no fault of his own, this child did not; and not unreasonably, applied it also to division of decimals. By this method 4.8 ∏ 0.6 came to 0.08. The same pupil had also learnt that if you know two angles of a triangle, you can find the third by adding the two given angles together and subtracting from 180˚. He got ten questions right this way (his teacher believed in plenty of practise), and went on to use the same method for finding the exterior angles. So he got the next five answers wrong." (Skemp, 1976, p. 23)

Fittingly, I believe Skemp's last sentence signals the problem perfectly: The boy got the next five answers wrong. Because he was blindly using the rule, the boy suffered from a severe lack of scope and sensitivity in his abilities. Memorisation and rule-following (further addressed in Section 3.2 and 3.3 respectively) are good examples of kludges, because they only provide a (very) small scope of abilities. Relying on a memorised answer or a simple algorithm may give you range or robustness, but not much scope or sensitivity, because your answers are only as good as what you've memorised or what you are able to calculate with the algorithm. Other examples are relying on cues of someone else or using a calculator.[87] Relying on cues from (or downright imitation of) someone else may give you scope and sensitivity, but it doesn't give you much range or efficiency because it requires the presence and cooperation of your partner. Relying on a calculator may provide efficiency, but you can only answer what the calculator can determine in the circumstances where you can use it.

Note that the concept of a kludge also draws attention to the fact that what distinguishes a system performing adequately and a system performing "kludgily" is a matter of degree, not kind. It is hard to call something a dirty shortcut if it does everything it is supposed to be doing. A shortcut that does everything appropriately is not dirty, but just efficient.

The kludge is an interesting problem, because it showcases my account's crux and virtue. It showcases a crux because it is about a way in which acts deceive, but it also showcases a virtue, because a kludge is still a deficiency of abilities, stability or system efficiency and should therefore, in principle, always be detectable. So no further marks (outside or behind acts) are necessary, even if further evaluations of further acts (beyond the limited set) may always be helpful.

Direct & Indirect Evidence

If a context of attribution values prediction, then an act of prediction is direct evidence of understanding because prediction is a valued and therefore a salient act. But not all salient signs of understanding need to be directly valuable and therefore direct evidence of understanding. For the purpose of evaluating understanding, I would like to draw a distinction between evidence that showcases understanding (i.e. the acts that comprise it) and evidence that merely implies understanding. I will call them direct and indirect evidence respectively. *Direct evidence* is any detection of an act that *comprises* the understanding and *indirect evidence* is any detection of an act that *implies* the presence of acts that comprise understanding. Consider an example: filling in missing

[87] Both imitation and the use of a calculator also bring with them problems of demarcating the appropriate subject that can be attributed with understanding (to be addressed in Section 3.3, but further developed in Chapter 4). But we can still address the problem of quality, regardless of whose quality it is.

gaps of a mathematical proof is a valuable act for the science of mathematics, and is therefore mathematically salient in understanding that proof. But now consider the ability of explaining something in one's own words. Even though being able to explain a proof-plan in one's own words may be valuable to particular mathematicians, it is not usually considered *mathematically* salient. It is therefore also not considered as a mathematical ability in itself. Nonetheless, being able to explain a proof-plan in one's own words is a reliable indicator of other abilities, such as being able to construct the proof (perhaps with minimal help), correct mistakes, find similar proofs, apply the plan to different situations, etc. Explaining something in one's own words is therefore good indirect evidence. So explaining in one's own words is a standard technique of evaluation not because one's own words are valuable to the practice, but because they are indirect evidence of abilities.[88] The difference between direct and indirect evidence is that the former is comprised of acts that are directly valuable whereas the latter is based on acts that have been discovered to be a reliable indicator of counterfactual (i.e. future or "possible") valuable acts.

Certain acts that serve as direct evidence may also be reliable indicators of further abilities. In an average subject, we may find patterns of certain acts being reliable indicators of generalised understanding, and then routinely apply those generalisations to apply in a specific subject. For instance: acts under pressure are usually a reliable indicator of range (presumably because if someone can perform under pressure, they can also perform without it). To be a reliable indicator, it needs to indicate other abilities, and it needs to do so reliably. If it fails to indicate other abilities, or fails to do so reliably, then its status of indirect evidence wavers along with it. This is, of course, in contrast to direct evidence. If predictions are directly valuable, then no act of prediction can be undermined as not being salient, on pain of inconsistency. They can be lucky and therefore misleading in scope, but what is undermined is the generalised understanding attribution, not the salience of the prediction act. By contrast, indirect indicators *can* meet with defeaters of salience, exactly because they are only as relevant as they are a reliable indicator of other valuable abilities.[89] If someone can successfully repeat or translate many proofs, but fails to display any other salient act, such as constructing the proof (or any like it), correcting mistakes, or any other mathematical ability, then this absence defeats

[88] This is also why I believe explaining in one's own words is not the ideal mark of understanding, even if it is ideal indirect evidence. Being able to explain in one's own words is a candidate mark I hear very often when I discuss abilities as understanding with lay people. Hills's (2009; 2015) list of appropriate abilities also included explaining in one's own words, although she was non-committent about whether these abilities *mark* or imply understanding.

[89] The distinction may not always be clear or relevant. For instance: is prediction only salient if its practically applicable or also if it's merely hypothetical? One could say that predicting hypotheticals is merely indirect evidence that the subject can handle real world predictions, because they are a reliable indicator of good prediction with real world data. But one could also say that the predictive aims of certain sciences go beyond our actual future. The point is not that every act needs to fit neatly into one category, but that we are able to express why certain acts can serve as evidence even if their salience as evidence may be defeated (because they themselves don't comprise understanding).

the relevance of the translation as an indicator of understanding.[90] Likewise, being able to explain the relativity theory or Gödel's incompleteness theorem to a layperson is a reliable indicator that one has understood the theory or theorem, but if one's abilities begin and end with a popularised explanation, then the scientific understanding is likely lacking rather than limited.

Implicit & Informal Context

The contextual approach opens up the possibility to express or make explicit from which context of attribution we are evaluating understanding. Making the entire context of attribution explicit would, however, be an arduous task. It may even be impossible to explicitly characterise an entire context of attribution in an exhaustive list of rigid and clearly delineated acts, circumstances and resources. Nonetheless, my contextual approach does not rely on there being such an exhaustive and finite list.

Mathematics is an interesting example here, because it is often supposed that it could be characterised entirely as a formal game of derivation, which entails that we could lay out most of its salient acts in an explicit, rigid and clearly delineated as well as finite way. But this characterisation has been heavily criticised by the philosophy of mathematical practice. (see e.g. Van Kerkhove, 2007; Mancosu, 2008) This is a recent movement in the philosophy of mathematics aiming to combat some of the misconceptions about mathematics with a focus on how it is actually practiced, so as to humanise our conception of it. Avigad (2008), who we encountered earlier, developed his account of understanding proofs with this in mind. Other scientific practices aren't usually any more clearly delineated in their salient acts or their interests. But the conversation about salience does not grow sterile in the absence of an explicit or clearly delineated context. With the critique on a formal derivation view, the philosophy of mathematical practice also directly or indirectly clarifies the true context of attribution. A relevant example here is automated theorem proving, which is a discussion which seems to most explicitly consider the context of attribution (even if it is not in the previously conceptualised terms). For instance: attention is called to the fact that computers excel at formal deduction, but that this alone does not put them on equal footing with human mathematicians. Formal derivations are not the locus of mathematical practice. But the discussion also called attention to the fact that the epistemic standing of both humans and artificial mathematicians can share similar worries regarding reliability. (Swart, 1980; Detlefsen & Luker, 1980) Nonetheless, contemporary artificial mathematicians do lack many of the abilities we value (and rely on) in human mathematicians. Indeed many of the things on Avigad's list (see Section 1.3) are absent in automated

[90] This may be why sometimes it makes sense to distinguish an understanding of the topic at hand from linguistic understanding of the topic's description.

theorem provers. For a closer look at how the informal practice of mathematics relates to the prospects of automation and artificial mathematical understanding, see (Delarivière & Van Kerkhove, 2017) and Chapter 6, Section 3. What is crucial about this discussion for our purposes is that that discussion still reveals the (contextual) salience of certain abilities, circumstances and resource-use.

So it was worth repeating that my account does not rely on being able to exhaustively characterise an entire context of attribution explicitly (i.e. by specifying all of the appropriate acts abilities it values, under which precise range of circumstances, and with which resources, along with thresholds for each) before one can assess understanding. What it does rely on is that cases of disagreement over understanding-attributions can be explained by pinpointing, indicating or recognising the nature of such a disagreement as either a difference in the context of attribution (e.g. one context values visualisability whereas another values abstraction more, or one context values prediction, whereas another values control), or as an inconsistency within one context of attribution (e.g. an astrologist may say they value prediction, while consistently favouring acts that appeal to an explanatory link with the position of the planets over acts of that appeal to predictions).

Dimensions & Kinds of Understanding

The variations allowed by a contextual focus also open up the possibility for carving up different kinds or types of understanding, even within the same practice. If it were relevant to do so, we can group certain uses, acts or abilities together (within or across different parameters) into one *kind* or *type* of understanding.

Imagine the following: Ro shows a vast scope of appropriate acts, but the circumstances where she does so are erratic. Tasha always acts at the appropriate times, but never with a great degree of sensitivity. Overall, a particular context may evaluate their quality as equal in degree, but their understanding is not equal in kind. The kinds are here distinguishable through their dimension (or parameter), but we can also distinguish kinds even within one dimension. For instance: imagine Nyota and Deanna have the same degree of scope, but where Nyota is good with numbers, Deanna is good in applying the models to real life cases. Even if their quality is roughly the same, the varying degrees of the parameters and variation within the content of these parameters allows us to specify what makes their understanding different. If the difference in their understanding is relevant, the context of attribution can be carved up to help us determine whether someone has a particular kind of understanding, meaning we can group certain uses, acts or abilities together (within or across different parameters) into different kinds or types of understanding (e.g. theoretical and practical

understanding). In other words, even within a single context of attribution, it can be agreed that Nyota and Deanna both understand, but that Nyota has a good theoretical understanding, but low practical understanding, whereas Deanna has a good practical understanding, but low theoretical understanding. If such concepts are useful, the parameters of quality proposed here can be used to focus on specific types of content and degrees to mark out a *kind* of understanding.

This form of contextual variation within a larger context of attribution can be exemplified in Skemp's (1976) famous distinction between instrumental and relational understanding. Instrumental understanding is limited to knowing a rule (being able to cite it) and knowing how to use it (producing correct responses that rely on straightforwardly using the rule). He distinguishes this from the broader relational understanding, which also includes why questions (and possibly more). Instrumental understanding is a widespread aim (even if it should not be the only aim) in teaching, and is easier to start with than relational understanding (which includes more abilities). A way of distinguishing instrumental from relational understanding is through which acts or abilities are salient to it. Instrumental understanding is satisfied with just the act of citing the rule and solving simple rule-based tasks, whereas relational understanding broadens the scope and sensitivity to include further appropriate acts (e.g. explaining the limits of the rule, answering why-questions, adapting the rule to the situation). It may be useful to wield a (sub)context of attribution that focuses on instrumental understanding (outside or along with a context of attribution focusing on a wider relational understanding) to detect a different kind of understanding or a different stage of development.

Another example of kinds of understanding is one which I mentioned much earlier. In Section 1.3, I explained that the literature on understanding makes a distinction between objectual understanding (i.e. understanding a topic, subject matter or body of information) from propositional understanding (i.e. understanding that something is the case), from atomistic understanding (i.e. understanding why something is the case). Based on such a distinction, debates can be had about whether objectual understanding and/or atomistic understanding can be built up from propositional understanding (or atomistic understanding) or not. But rather than use propositional understanding (and all its problems - see 1.2 and 1.4) as a building block, my conception of understanding builds from acts and abilities (along with its presence across and through circumstances, and the resources used to achieve them). A way of distinguishing objectual from propositional from atomistic understanding could be achieved with the salient type of acts and abilities within the scope (and sensitivity) parameters. Furthermore, because kinds of understanding, under my approach, are merely conceptual categories to help us specify the salient content and thresholds within contextual parameters, the problems of overlap or

reducibility between objectual, propositional and atomistic understanding doesn't pose such a problem as it does now.

On a related note, there's a consequence of this contextual account not just for kinds of understanding, but also for *kinds* of explanation. This approach may clarify why there's a plurality of accounts about explanation, and a struggle to find a single account that exhaustively covers what feature(s) makes something explanatory. If we accept that an argument is explanatory because it grants understanding (see Wilkenfeld, 2013b, but also Delarivière, Frans & Van Kerkhove, 2017), then different accounts of explanation - for example Kitcher's (1989) unification, Steiner's (1978) characterising property, or Lange's (2014) salient features - all target a type of argument that grants certain abilities within scope and sensitivity. But they may simply target explanations that grant different abilities within that scope or sensitivity (with some overlap). Even if we may ultimately find a single account that covers all forms of explanatoriness (that covers granting the entire salient scope and sensitivity of understanding), it seems that at present, accepting pluralism is heuristically more fruitful (see Delarivière, Frans & Van Kerkhove, 2017). Furthermore, the contextual account presented here seems to suggest that the content within scope and sensitivity of understanding is altogether too large and too varied in salience to expect to find such a single account.

In sum, if a practice would find it useful to do so, any explicit or implicit, rigid or vague set of acts or abilities could be grouped as a relevant kind of scope, situational responsiveness, accuracy, range, robustness, economy or potential, or combination of the aforementioned. This can then be used to discern different *kinds* of understanding.[91] Once again, I am not implying all kinds require an explicit and exhaustive list of which acts, abilities, range, robustness, economy and/or potential is required. What I am saying is that their differences could be marked out as differences in the salience within these parameters and that doing so may help us prevent talking past one another.

Maximal & Minimal Understanding

Given the ability-approach to understanding and the contextual approach to the salient abilities, it is now pretty straightforward to see what the lower bound of understanding would be: no understanding means no appropriate abilities at all. What about the upper bound then? Is there such

[91] Sierpinska (1994) calls these "ways of understanding" (p. 4). I prefer to use the word kinds because "ways" implies that we are talking about a particular process of implementation, or mental process, which would be speculative at best. Skemp (1976)'s distinction has a similar problem. Because his conceptualisation is not act-based, relational understanding isn't just distinguished from instrumental by distinguishing the abilities present, but also in the supposed mental schemes that lie behind them. The dangers of this were discussed extensively in Chapter 1, so I won't repeat them here.

a thing as *complete understanding*? In practice, such a thing seems unlikely, but my approach should be able to give a sense of what complete understanding would entail in principle.

The concept of complete understanding is not often discussed. An exception is Kelp (2015), who defines "maximal understanding" as having "fully comprehensive and maximally well-connected knowledge" (p. 17) of the phenomenon in question. In defining "outright understanding," he expands that with "such that S would (be sufficiently likely to) successfully perform any task concerning P determined by c, if, in addition, S were to have the skills needed to do so and to exercise them in suitably favourable conditions."[92] Under my presented approach, we can make a similar claim, namely that maximal or complete understanding is attributable with all of the salient abilities in all the salient circumstances and with only the admissible resources. If there is only a limited scope, stability and system efficiency that is of contextual salience, then it should be possible, in principle, to have a complete understanding, given that those limited demands are satisfied. Although I expect very few contexts of attribution to be sufficiently limited to be able to acquire complete understanding in practice, it can be regarded as a strength of my account that it does allow us to speak of complete understanding in those rare cases where our epistemic aims are limited enough. Furthermore, for the same reason, my account also shows why complete understanding will usually be unattainable: our epistemic aims are ever shifting and ever growing (thanks in part to recursion). No sooner is one aim satisfied, than we can think of a new one. Our epistemic aims tend to exceed our epistemic grasp.

In Sum

In this chapter, I conceptualised the dimensions and degrees of quality in understanding, offered up a contextual approach to specifying what is salient, and specified some of the problems and opportunities in evaluating understanding under that approach.

It was largely agreed that understanding is not binary, but comes in levels or degrees. Understanding, unlike knowledge, requires an expression of not just its presence, but its quality. So, instead of supplying the necessary and sufficient conditions for understanding (which leads to problems even for conceptualising knowledge), I approached understanding and its quality as carved up of degrees and dimensions. Unfortunately, and quite unsurprisingly, no agreed single universal standard can clarify all attributions of understanding within these dimensions. We found contextual variance not only in

[92] Unfortunately, this does not seem to include the stability dimension (meaning he can't distinguish the difference in degree of understanding between, for instance, two subjects with the same scope, but a different degree of success in range). Something like stability may be implicit in "sufficiently likely" and "suitably favourable conditions," but this depends on what the modifiers" "suitable" and "favourable" are doing exactly.

thresholds, but in what is considered salient in the first place. Therefore, I also offered up a contextual approach to each of the dimensions and parameters, allowing each of them to vary what is appropriate (while leaving the justification of what is appropriate to another discussion[93]). We saw four dimensions of quality, three of which were composed of two parameters (one which widens it and another which deepens it).

The first dimension was that of the *scope* of abilities*,* which tracked the amount of variety in abilities salient to understanding an object X (which could be conceptualised through its meaning - see Section 1.4). For instance: you display your understanding of the theorem not just by supplying a proof, but in giving a rough outline of the proof, supplying different proofs, using the theorem where it is appropriate to do so, showing what would happen if the theorem were false, etc. Nonetheless, which abilities are salient to understanding object X and how many of them would suffice needn't be agreed upon by all parties interested in attributing understanding of object X. *Scope* or *domain weights* were a way to conceptualise which uses connected to X are deemed appropriate, and to what extent. Each domain could be distinguished or grouped by its field of problems, intended aims, empirical standards, conceptual tools, methodological constraints, logical requirements, standardised symbols, background metaphysics, field of focus, or pedagogical stage. These will dictate which abilities are salient to that domain and for its attributions of understanding. The display of each of them makes an understanding attribution more warranted, but none of them are necessary or sufficient by themselves.

The second dimension was the *sensitivity* of an ability. The sensitivity parameters were *situational responsiveness* (i.e. amount of appropriate changes in performance to changes in the object-situation, e.g. responding to what-ifs) and *accuracy* (i.e. degree of precision in performance, e.g. number of decimal points). Nonetheless, both the content and threshold of the sensitivity parameters needn't be agreed upon by all parties interested in attributing understanding of object X. *Situational* or *what-if weights* are a way to express which variations in object-situations, along with their appropriate reactions, are relevant, and to what extent (for a context of attribution). For example, when it comes to understanding space-trajectories, the what-if of an additional planet might be more salient to NASA, whereas the what-if of an additional cosmological constant might be more salient to theoretical physicists. Next, *accuracy weights* are a way to express which types of accuracy are appropriate (when there are degrees to success) and to what extent. For example, what degree of accuracy is salient for

[93] Nonetheless, I did make some remarks about what makes a context of attribution justifiable scientific as well as why the implicit assumption of neurotypicals should not imply neuronormativity.

understanding the position of the sun is different for a navigator than it is for a physicist. And what type of accuracy (e.g. long term versus short term predictions) is salient for understanding the weather is different for a forecaster than it is for a climatologist.

The third dimension was the *stability* of an act. The components of the stability parameters were *range* (i.e. degree of presence in (counter)factual circumstances) and *robustness* (i.e. degree of presence after (counter)factual circumstances). For instance: the ability to produce a proof is stable if it can be carried out by our same subject regardless of variations in the circumstances which she finds herself in (e.g. weather, time of day or location) or having gone through (e.g. confronted with misleading information). By this I don't of course mean that she needs to do so under or through *all* circumstances, but in as many salient circumstances as possible. *Range* or *deployment weights* are a way to express which types of (counter)factual circumstances, where the same (barring non-epistemic changes) subject acts appropriately, are salient and to what extent (for a context of attribution). There's a time and a place for everything, and here we can specify what they are, contextually (e.g. Vir, as opposed to Londo, displays his abilities in the appropriate circumstances for academic purposes, but not social ones, and vice versa). *Robustness or rationality weight*s are a way to express which types of circumstances, after which the subject needs to continue to act appropriately, are salient and to what extent. This can, for example, also specify which types of information should not (and should) easily sway our subject (i.e. what is good and what is bad evidence).

The fourth and last parameter was the *system efficiency* of a subject, composed of the economy parameter (where the appropriate act uses a minimum of saliently allowable resources) and the potential parameter (where the appropriate act obtains with the addition of a minimum of salient resources or events). While separating a subject from certain external resources can be telling about her abilities, scientists don't usually work in a vacuum, so the evaluation of their performance should be able to include all of the used resources. *Economy* or *resource weights* were a way to express which particular resources (incl. events) one is willing to allow to be used to consider the abilities achieved, and to what extent (for a context of attribution). This is different for secondary students than it is for research mathematicians. Next, *potential weights*, were a way to express which particular resources and circumstances one is willing to consider in order to assess the abilities achieved with it, and to what extent. Usually, this will be tied to the socio-economic costs and possibilities thereof.

These dimensions and parameters are useful conceptual tools. Firstly, they allow us to express a quality of understanding. They don't give us a tool to measure the precise quantity of understanding,

but they do give us a tool to help us clearly indicate which things warrant a better understanding and which things can undermine it, and so they can help us decide what to look for and where. Even though most authors acknowledge the degrees of understanding, few of them address them as explicitly as was done here. Secondly, they allow us to pinpoint different kinds of understanding depending on how the subject fares in each parameter. Thirdly, they dissolve the need for certain conditions (e.g. anti-luck conditions) which have proven unwieldy. Fourthly, they provide a basis for contextual variations in understanding attributions. And lastly, my account can easily incorporate something akin to all-or-nothing attributions by using a threshold. It is my contention that most attributions of understanding will boil down to a claim about the degree within these dimensions. Even if these parameters are imperfect in conceptualising the "ideal" assessments of the quality of understanding, they are fruitful in diagnosing the strengths and weaknesses in quality as well as the problems in evaluation - as will be attested by how well they fare in addressing tricky or misleading attributions and proposed counterexamples to the ability account (the topic of Chapter 3).

To end, I conceptualised the evaluation side of understanding by addressing some of its aspects. I labelled a common misevaluation as involving a kludge (a dirty shortcut leading to a limited set of successful performances that forces us to overgeneralise if they are the ones we happen to be witnessing). Nevertheless, such kludges can still be sniffed through their inappropriate acts (which is what makes them "dirty"). I introduced a distinction between direct and indirect evidence (namely, the evidence that showcases the acts that comprise understanding, and the evidence that merely implies such acts will also be present). I pointed to some limits of characterising a context of attribution (namely, the recursive and informal nature of some practices), but also that this doesn't stop us from indicating at the context of attribution. I also showed how contexts of attribution can handle kinds of understanding (by further carving up a context of attribution according to a precise focus in saliences), and considered what it might mean to have complete understanding (namely display the full scope, sensitivity, stability of acts with only the admissible resources) even if this seems, in practice, to be well-nigh impossible.

PRELUDE 3

The Illusion, Quality and Tutor of Discretion

DIRECTOR: And, scene! Thank you. That was a good run, given the hiccup there. Some very good improvisation, Rose. Very clever of you to add the example of "variations depending on the axioms." That's just what Rosencrantz would have said.

ROSE: Thank you. I do still struggle with those last lines though.

DIRECTOR: That's okay, we'll make those lines rhyme so they are easier for you to remember.

ROSE: Thanks, that would help me.

DIRECTOR: And good of you, Gill, to then continue the scene using that example of the axioms and postulates. Exactly what Guildenstern would have said. Especially since you got it wrong. I thought that was funny.

GILL: I would never have thought about it if it weren't for Rose.

ROSE: Oh, but your improvisation also helped me improvise further.

GILL: At any rate, I'm glad you liked it.

DIRECTOR: I did, I did. Very appropriate. Now let me check my notes. *(Beat)* Okay. Esther, you also did well as the Professor. You're doing excellent on the lines. However, you do seem to be struggling with the actions. I thought of a solution though. I'll ask the writer to add the actions to your lines and then you'll remember what to do because you're saying what you're doing.

ESTHER: That's good, thank you. Would it also be possible to have some scene-related props? They always help me as well.

DIRECTOR: I'll speak with the prop master. Consider it done.

(Pause)

DIRECTOR: Now. Hamlet.

KENNETH: I'm sorry.

DIRECTOR: What went wrong? You were supposed to come back at the end of the scene for a final speech, but you didn't.

KENNETH: I know, but as soon as Rosencrantz & Guildenstern started changing their lines, I no longer had my cue and I just didn't know what to do.

DIRECTOR: I want you to be *Hamlet*, the erudite speechmaker.

KENNETH: You are aware that I am not *really* an erudite speechmaker?

DIRECTOR: I know that, but what I want you to do is to use your acting skills to *portray* one.

KENNETH: How do I know what to say if the words aren't written down for me in the script? How do I know where to stand if no one tells me?

DIRECTOR: Let your own discretion be your tutor.

KENNETH: That's just it. I don't have the appropriate discretion to tutor me.

DIRECTOR: Okay, well, we'll have to fix a way to uphold the illusion that you do.

KENNETH: What if you whisper the appropriate responses to me?

DIRECTOR: I can't just control every detail of this production. I'm afraid *my* discretion just isn't big enough. We each have to do our part in working together as a team.

KENNETH: Well. I *can* pretend to be an erudite speechmaker if I have the lines memorised, so why don't we come up with a few canned responses for a couple of eventualities?

DIRECTOR: Oh, Kenneth, You're such a klutz. There are so many eventualities, and there'll only be a select few of them where we can get away with a canned response.

KENNETH: So what if we just prepare a text for *every* possible eventuality?

DIRECTOR: Right, where are those monkeys with their typewriters? Never mind, we don't have the time. We open in two weeks, Kenneth. *(Beat)* We'll just have to hope that nothing veers off script.

Rehearsals for the next two weeks went surprisingly well. The props helped Esther remember her actions, the rhymes helped Rose & Gill remember their lines. The Director could focus on only correcting courses instead of coming up with a plan for every detail. And when mistakes were made, all of them managed to correct course so Kenneth could stick to his pre-appointed script.

Unfortunately, on opening night, it became clear that Rose & Gill both had stage-fright, so neither ever uttered a single word, Kenneth was thrown off balance and Esther performed her actions even though it was no longer appropriate to do so.

Chapter 3

ADDRESSING OBJECTIONS

If abilities are the true mark of understanding, as I have argued in Chapter 1, then finding counterexamples that showcase we can have understanding without abilities or abilities without understanding would undermine that approach. Wilkenfeld (2013a) certainly believes an anti-ability claim could be substantiated in this way:

> "the difference between understanders and non-understanders is that the former, but not the latter, can utilize the understood effectively. But of course various factors prevent or empower one to affect things in the world without being signs of understanding or its absence. Inability to do derivations in first-order logic could arise from a broken pencil; conversely, ability to perform such a derivation could result from having memorized this particular derivation or just being extremely lucky." (Wilkenfeld, 2013a, p. 1003)

He already presents a few suggestions here, such as the lack of appropriate tools, memorising answers and luck. We will now consider these, as well as a series of other candidate counterexamples (some of which I have already touched upon briefly), and I will show why each of them fails to hurt the ability account as presented here. In doing this, I will further validate my approach to understanding discussed in the previous chapters, and showcase how it deals with many of the staple examples to be found in a variety of literatures.

First I will cover, in Section 3.1, those candidate counterexamples that seem to warrant an understanding attribution, but where abilities seem to be lacking. These involve cases where abilities (i) are masked, (ii) lie outside of non-standard circumstances, (iii) are deliberately avoided, (iv) are (temporarily) impaired, (v) are finked, (vi) would require tools, (vii) are lacking due to low technical skills or (vii) bad luck. I will try to show that the presence (or absence) of understanding in each of these cases can be recast as direct or indirect claims about the (counterfactual) scope, sensitivity and stability of the (salient) abilities - thereby keeping abilities in their role as the mark of understanding.

Next, I will cover the candidate counterexamples that seem to involve abilities, but where the understanding attribution seems unwarranted. I believe these can be roughly divided into two types,

depending on why they fail to counter the ability approach as presented here. The first type of examples, addressed in Section 3.2, are examples where the abilities are due to (i) a lucky shot, (ii) environmental or evidential luck, (iii) gettier luck, (iv) rote memorisation, (v) false beliefs, theories or idealisations, (vi) a short term, (vii) employing algorithms or models, or are merely (viii) emulated. I will argue that each of these candidate counterexamples fails to counter my account because they are still trying to warrant understanding through the lack of counterfactual acts, thereby failing to show that the appropriate abilities are indeed present (or indeed lacking).

The second kind of counterexamples that focus on abilities-without-understanding will be addressed in Section 3.3. Here, I will cover those examples where the abilities are due to (i) mimicking, (ii) reverse finks, (iii) external resources, (iv) a giant look-up table, or (v) blind rule-following. I will argue that for each of these candidate counterexamples, the failure to counter my account comes from attributing the understanding or abilities to the wrong subject. To end, I'll also discuss abilities that are (vii) derived from others (where it is mistakenly presumed that the understanding is therefore attributed to the wrong subject), and (viii) lacking in coherence (where it is mistakenly presumed that the lack of coherence needs to be conceptualised and addressed in the mark of understanding or the subject with understanding, rather than through the object of understanding).

3.1 Understanding without Abilities Objections

If abilities are the true mark of understanding, as I have argued in Chapter 1, then it would need to be impossible to find a counterexample where we can attribute understanding even in the absence of the appropriate abilities. When presented with such an example, I would either need to show why that case is not inconsistent with the ability approach, or remove abilities from their throne as the mark of understanding. However, to present such an example would require a motivation for the validity of the understanding attribution that is not reliant on soft intuitions (i.e. weak evidence that may be incorrect and may vary from person to person), overly speculative psychology (i.e. based on another equally unsubstantiated claim) or incoherent metaphysics (i.e. varying the base of a claim based on bias), and we've already seen how this can be a difficult task. The difficulty of that task becomes all the more clear under the counterfactual account of abilities - because you would need to motivate the understanding attribution *without* the presence of appropriate acts, even in the relevant counterfactual worlds. In this subsection, I will be addressing counterexamples to the ability account that (seemingly) showcase a lack of abilities, but where we have a good reason to believe there is understanding. I will try to show that each one of these can be recast as direct or indirect claims about abilities, after all.

(i) Masks

The first kind of counterexample that I'd like to consider is that of *masks* or antidotes. They are particularly relevant because masks or antidotes are standard counterexamples for dispositional accounts, and my conception of abilities sails pretty close to the concept of dispositions. What is a disposition then? The simplest analysis of a disposition is fittingly called the *simple conditional analysis*. The gist of it is captured in this example: Glass is *disposed* to break when struck, if and only if, it breaks when struck. Although the simple conditional analysis seems intuitive, there are many counterexamples that form a problem for it, such as those involving masks (which we'll deal with here) and finks (which we'll get to later in subsection v). The former kind involves a "mask" (Johnston, 1992) or "antidote" (Bird, 1998). For instance, if glass is protected by packaging material, then we can say that the disposition to break is masked by the packaging or that the packaging provides the antidote to glass breaking. The mask or antidote has "the effect of breaking the causal chain leading to [the response], so that [the response] does not in fact occur." (Bird, 1998, p. 228)[94]

In like spirit, we may say that a subject's abilities can be masked. These masks can be externally applied (e.g. a racketeer saying "don't even think about doing anything or I'll kill you" to the subject) or intrinsic to the subject (e.g. the subject being incredibly shy with a large group of people around). The barrier between intrinsic and external is not always clear-cut, but this does not change the argument (see the next paragraph). What's key about cases of masks is that they are, by their very design, not cases where abilities are ubiquitously *lacking*, but where they are *prevented from obtaining* (be it from within, or without). Counterfactuals can distinguish the two by revealing the presence of abilities where the mask is absent (e.g. in the absence of the racketeer or the perception of large crowds). The mask is the difference-maker, but the relevant difference it reveals is the presence of abilities.

Furthermore, if we can find a pattern of difference-makers (e.g. racketeers, a large crowd), then it is perfectly possible for us to contextually devalue the range that involves that difference-maker as especially salient (just keep her away from racketeers and large crowds), as opposed to other types of difference-makers (e.g. just give her weeks to respond, along with a group of experts, the internet and a calculator). If we can identify masks, we can decide whether the range of circumstances that exclude

[94] Antidotes or masks are counterexamples to the simple conditional analysis, but were actually intended to counter a more robust analysis, namely that of Lewis's (1997), which has the additional requirement that the disposed subject/object must also have (and retain) an intrinsic property which can serve as the causal basis for its disposition. This protects the analysis from finkish dispositions (see later), but not from antidotes, which affect the causal basis itself. (Bird, 1998) If a glass fails to break when struck, this is not because it isn't really struck, or because its causal basis for breaking has disappeared, but because of the presence of the antidote: the packaging. The antidote or mask works in such a way that the disposition "is not manifested even when the appropriate stimulus conditions are present and the causal basis remains intact." (Choi, 2011, p. 1160)

them are salient to consider (e.g. is it relevant to understanding that the subject is surrounded by a large group of people or not? Is it relevant that the subject gets death-threats?) or whether excluding them from circumstances is worth doing (e.g. Is it too much to isolate the subject from large groups? Is it too much to prevent someone from making death-threats?). Using counterfactuals doesn't just draw open the space of possibilities, it also allows us to detect, consider and assess difference-makers. This is not to say masks are always easy to detect, but the mere concept of masking that this counterexample relies on does entail that we have a means of pointing to certain difference-makers *as a mask*. Furthermore, because the example relies on the *prevention of abilities*, abilities are still doing their job as the mark of understanding. The crux of masking is in the difficulty of evaluation, and in a particular limit of circumstances (be they salient or otherwise), not the absence or irrelevance of abilities. Therefore, what marks the subject's masked understanding is not to be found beyond its abilities, but beyond its mask.

An important real-life example of masked abilities, which showcases the vagueness of the intrinsic versus external distinction (because it doesn't neatly fit in either the external or intrinsic category), as well as the subtlety of detecting the difference-maker (because it is not hidden, but it can be easy to miss) can be found in a specific case of internalised oppression. Internalised oppression is a situation where the subject "comes to use against itself the methods of the oppressor" ("Internalized oppression," 2019). One such form of internalised oppresion is internalised sexism, where a person incorporates learned sexist behaviors and attitudes, even towards themselves and people of her/his/their own sex or gender. If the behaviour or attitude is about the lack of competence of the group in question, then a lack of confidence or impoverished self-awareness about one's own competence can lead to actual (or potential[95]) abilities being masked. For instance:

> "[D]ue to sexism, girls are provided with few female role models in the sciences and may meet with low expectations or discouragement on the part of adults about their mathematical and scientific abilities, even about their intelligence in general (Eccles, Barber, Jozefowicz, Malenchuk, & Vida, 1999; Jacobs, Davis-Kean, Bleeker, Eccles, & Malanchuk, 2005). A girl in such conditions may internalize the inequity and declare that she's just no good at math and science. Feelings of powerlessness may be expressed as assertions of incompetence which may in turn reinforce the sense of powerlessness and powerless behavior." (Bearman et al, 2009, p. 15)

[95] See the potential parameter in Section 2.1.

The low expectations and discouragement, as well as the resulting feelings of powerlessness and diminished belief in their own abilities can *mask* the actual (as well as potential) abilities in female students. But when we determine that these students still warrant understanding attributions, it is not *in spite* of abilities, but because of them.

(ii) Non-Standard Circumstances

Imagine Jane is a graduated mathematician applying for a teaching job and we just want to have a quick check whether she is up for the task. We ask her to explain to us why the square root of two is irrational. What we would like to see and/or hear from her is a proof, some clarifications, critiques of our proposed missteps, some clarifications on what the misstep would lead to, etc. Yet she does not perform any of these actions. Instead, Jane keeps yelling "Aaaah, it's so hot!". I should perhaps have mentioned that the interview was being held on top of a volcano, which we can safely call *non-standard circumstances.* Should the lack of acts make us withhold an understanding attribution? And if not, why would we be justified in attributing understanding in spite of the lacking acts?

The ability-approach does not entail that we need to read our evidence so narrowly that attributions are only warranted in the precise range of circumstances where the subject is acting (or not acting). This disregards the modal reading of the concept of ability (see Section 1.4). Of course, the approach also doesn't require of us so broad a reading that any circumstances where there is an appropriate act would vindicate the attribution, no matter what the difference-maker is or how many are necessary for the act to obtain. So it is worthwhile to note that in standard circumstances (e.g. in an office with room-temperature) Jane does display all the appropriate acts. And if she didn't, it would be more difficult to argue why she nonetheless warrants an understanding attribution even as she fails to act while dangling over the volcano.

Even if we don't want to specify what precisely the standard circumstances are or should be, finding a difference-maker will be relevant. There is a sense in which the volcano masks Jane's ability. Furthermore, if we can find a pattern of difference-makers (e.g. being subject to temperatures higher than 35°C), then it is perfectly possible for us to contextually devalue (or diminish) the salience of the range that includes this difference-maker (e.g. only consider the subject in circumstances roughly around room-temperature), as opposed to others (e.g. the subject is provided with a cheat-sheet). In one sense, one could say that abilities are superior if they include non-standard circumstances (in that they have a broader range of circumstances under which they can be displayed - including on top of a

volcano, and perhaps other circumstances of physical discomfort), but in another sense, we are keen to dismiss the relevance of acting under non-standard circumstances.[96]

Whichever avenue we pick (be it drawing open the relevant counterfactuals, or denying the salience of unfair circumstances), the understanding in these cases are vindicated by their abilities in the salient circumstances.

(iii) Deliberate Avoidance

Imagine there is an expert who simply denies to act appropriately. Ben has worked for many years as a successful trader for a large bank. Over many years, Ben has come to understand a lot about how Wall Street works, and he is fed up and no longer wants anything to do with it. So he decides to *deliberately avoid* responding. Each of our probes into his abilities, be they theoretical (asking for explanations, clarifications, mechanisms, etc) or practical (asking to navigate in the system, securing an ISDA, etc) fail to result in the acts appropriate for our understanding attributions. So does Ben still understand anything about Wall Street? This is what Fara (2008) calls a "straightforward failure to exercise the ability" (p. 846), and it is something everyone routinely does. "I have the ability to smash all the windows in my house, but I routinely fail to exercise that ability. I fail, on these occasions, because I do not even try" (Fara, 2008, p. 846). Even though Ben seems to justify an understanding attribution, he fails to show any of the abilities that would warrant one. So what is going on?

This case is not inherently different from any we have seen before. This is just another case of masking, except the masking is more intrinsic (i.e. conceptually connected to the interests of the subject - more on that in Chapter 4). It is relevant to note that the mask (i.e. his disdain for Wall Street) seems to be the main difference-maker. We can ascertain that Ben used to have abilities, and the only relevant difference is that he no longer wants to act in the way appropriate for understanding attributions. The relevant counterfactuals to reveal Ben's mask are the ones where his motivation to refrain is trumped (because someone close to him requires his services, because his life depends on it, or because he is convinced that Wall Street is not evil after all, etc). This case is relevantly distinct from a case where Ben has amnesia and no longer remembers anything, where the counterfactuals under which Ben

[96] In standard circumstances (in an office with room-temperature) Jane displays all the appropriate acts. So does Jean. Jean, however, has a genetic mutation and has heat-resistant skin, so even on top of the volcano, she displays all of the appropriate abilities. Does Jean understand more than Jane does? In one sense, all other things being equal, one could say that Jean's abilities are superior (in that they have a broader range of circumstances under which they can be displayed which includes on top of a volcano) but in another sense, we may find the salience of acting under non-standard circumstances too low to consider (see how contextual interests can vary in Section 2.2).

would respond appropriately again would need to include going through years of study and experience again. In the latter case, Ben truly has no abilities, and has therefore truly lost understanding.

Of course, if it is possible that the appropriate acts can be avoided in actual circumstances, then there is an implication that it is theoretically possible for someone to become an expert without *ever* displaying evidence for the appropriate abilities in actual circumstances. Nonetheless, an attribution of understanding even in the absence of actual acts could not get off the ground without invoking appropriate acts in the relevant counterfactual worlds. Of course, it is true that an attribution based on counterfactuals alone would always be highly speculative. An approach to understanding that avoids abilities as a mark, however, is in no better state to deal with such a case. For instance: mental state approaches can not account for Ben's understanding any better than ability approaches do. Until there are mind reading techniques (that do not rely on external cues), their guess about mental states is no better than ours about counterfactual acts. The best thing we can say with any confidence about extreme cases like these is that it is unclear, due to evaluative limitations, whether that subject does indeed understand. And would we want it any other way?

(iv) Impairment

Not all things that act like masks are equally salient difference-makers. To explore this, I'd like to consider the notion of *impairment.* Chomsky (1997) has used an impairment-example to attack the idea of marking something through abilities. The case he makes is against knowledge, but the argument, if sound, would apply equally well to understanding. His claim is that the "ability to use language can be impaired, and can even disappear with no loss of knowledge of language at all." (p. 13). To illustrate, he presents us with the following:

> "Suppose that a speaker of English suffers Parkinson's disease, losing entirely the ability to speak [and doing all the things associated with using English] and therefore does not have knowledge of English, as the term is defined by [the ability account]. Suppose that use of the chemical L-Dopa can restore entirely the person's ability, as has been claimed (it does not matter whether the facts just noted are accurate; since we are dealing with a conceptual question, it is enough that they could be, as is certainly the case). Now what has happened during the recovery of the ability? On the assumption in question, the person has recovered knowledge of English from scratch with a drug, after having totally lost that knowledge. (...) Had the person been a speaker of Japanese, [s]he would have recovered Japanese with the same drug. Evidently, something remained fully intact while

> the ability was totally lost. (...) [This shows] that knowledge cannot be reduced to ability." (Chomsky, 1997, p. 13).

We can easily substitute "knowledge" for "understanding" here, so I will address Chomky's claims about knowledge as claims about understanding. Chomsky claims that, because the abilities were regained with a relatively simple solution and without *relearning* them in the usual way, knowledge (or understanding) must have been retained even in the absence of those abilities.

> "Note that there are cases where we would say that a person retains an ability but is incapable of exercising it, say, a swimmer who cannot swim because [her] legs and arms are tied. But that is surely an entirely different kind of case than the one we are now considering, where the ability is lost, but the knowledge is retained" (Chomsky, 1997, p. 13)

Notice, though, that this argument still relies on the speaker of English to *recover* abilities as opposed to *relearn* them. Whatever was "retained" throughout the onset of Parkinson's is revealed after the administering of the drug. And what was revealed are abilities. As such, one could say her abilities were masked (the difference-maker being rooted in Parkinson's and removed by L-Dopa). If the speaker of English could not recover, except through learning English a-new, then there was no ability to mask, nothing to retain that warrants the knowledge or understanding attribution. The understanding attribution is warranted only because we can establish that the speaker can recover abilities, not because we are establishing something beyond abilities. One could insist that what it is really about is not establishing abilities, but the causal basis for them. But the ability conception does not entail denying a causal basis for the ability. What it does entail is that, only to the extent that that basis delivers abilities, can it mark understanding. So if we remove a mask that messes with that causal basis, only the causal mechanisms that lead to abilities would vindicate the attribution. That is why recovery is different from relearning.

Nonetheless, no one can deny that the abilities were masked to an unusually severe degree. And the severity here is relevant, both for the abilities and the understanding. To say that the patient keeps the abilities because there is *something* which can bring them out is perhaps a bit of a stretch. Chomsky anticipates this. He goes on to say that only on a far-fetched philosopher's notion of "ability", which he calls P-ability, would we say that this person suffering from Parkinson's disease retains her ability (in the P-ability sense) to speak English, even though she has lost her ability (in the normal

sense) to speak English. So "nothing has been achieved except that we now mislead ourselves into believing that we have maintained the thesis that knowledge is ability". (Chomsky, 1997, p. 14). I concede that I have no interest in broadening the notion of ability wide enough to capture such cases, but the reason why is not because there is something beyond the abilities, but because the difference-maker is so extreme (far removed from normal or salient circumstances) that it is a stretch to call it an ability.

Nonetheless, saying that the patient retains her understanding is also a bit of a stretch. The severity and ubiquitousness of the mask is what would stop me from attributing understanding - the circumstances with the mask are (for most contexts) more prevalent and more salient than the ones without it. In those contexts, the most one can say is that there is a *potential* for the patient to regain her abilities and understanding. And an above average potential at that, because it won't rely on the patient re-learning the language, but instead on the patient taking a particular type of drug. If this drug would work effectively and be readily available, conceptualising Parkinson's as a mask would become more salient, but unfortunately, the effects and availability of the drug is a stretch. Therefore, its salience in our current world is not that high. For the sake of argument though, let's imagine if it wasn't such a stretch. Let's turn the knobs of the thought-experiment[97] and say that both Parkinson's disease and L-Dopa are as commonplace as drowsiness and coffee. We don't deny someone with understanding because they need a coffee to wake up from drowsiness before they can get to work. But that is because, for most contexts, the difference-maker of coffee is easy and ubiquitous. If drowsiness was more severe as well as lasting, and coffee more expensive as well as rare, is there any difference between the two? Conversely, if the cure for Parkinson's were as commonplace as coffee and waking up, wouldn't treating it as a mere mask become particularly salient? And, more importantly, if there was no possible fix (no coffee, no L-Dopa) and therefore no abilities to quickly unlock (except through the same arduous process of learning it the first time), would we have any grounds for thinking that something was "retained" which warrants the understanding attribution? If you think we do, the burden of proof is on you to argue what it is.

Now, what about *permanent impairments*? A classic example is that a pianist without hands still knows how to play the piano even though she has lost the ability to do so. (see Stanley & Williamson, 2001) The same argument can be made for understanding. Imagine Luke used to be an A-grade pianist who recently suffered a terrible accident that made her lose both hands. It may be fair to say that she still understands how to play the piano. Nonetheless, she won't physically be able to play the Moonlight

[97] A tactic for exploring thought experiments taken from Hofstadter (Dennett & Hofstadter, 1985, p. 375)

Sonata, or indeed any other piece. So isn't this a case of understanding without abilities? I contend that this is either an extreme (and therefore usually not salient) claim about her ability, or simply about her other abilities. I'll elaborate on both.

Firstly, it could be read as an extreme case of masking. I would go so far as saying that it is so extreme, that the difference-maker (i.e. getting new hands) is only salient in very speculative contexts. So the counterfactual claim is that if Luke were to have never lost her hands, or given new synthetic ones, she would play as beautifully as before. But because we can't simply give Luke any new hands, this at best constitutes an interesting speculative potential. Nevertheless, it is a potential not usually found in people who can't play - most people, whether they had hands to begin with or not, would not be able to play the piano even when they have hands or get supplied hands when they don't. That does mean that the difference-maker is radically different from a difference-maker in most novices (e.g. years of training), but these circumstances are not salient enough *for most practical purposes* such that we can stand by the claim that she is still able to play the piano. If one insists that her understanding is unchanged, one has to rely on a context of attribution that recognises the salience of the outlandish counterfactual, and if one insists that she is not able to play, one has to deny the salience of the outlandish counterfactual. While both contexts of attribution are valid, it would be inconsistent to jump from one to the other within the same analysis.

Secondly, even if we don't allow the salience of the outlandish counterfactual (where she still has hands), it would still sound wrong to say that Luke now no longer understands how to play the piano *at all*. So is it a case of understanding without abilities after all? No, not without abilities, plural. Understanding how to play the piano involves a scope (see Section 2.1) that is a lot wider than merely being able to play pieces on it. She can explain how to play to someone else, she can recognise difficult moves, detect mistakes and explain why they happened as well as how they should be avoided, etc. In this sense, the counterexample is even more difficult to motivate if it involves an understanding something that is not so highly reliant on bodily movement (e.g. mathematics).

So there is a sense in which the pianist can be said to play particular pieces, but only in that (a) the relevant difference-maker would be to still have (or get) hands, (which for some contexts of attribution can be relevant) and (b) there is a sense in which the pianist understands how to play the piano that doesn't involve the ability to actually play it, namely the wide scope of salient abilities beyond it (can explain how to, can teach it, can criticize others for playing incorrectly, etc). What is crucial here is that understanding something will typically involve a lot of appropriate acts across many salient

circumstances and that what warrants the understanding attribution in these objections is rooted in the same modal claims about abilities as those of the ability-account, so there is no shift in what marks the understanding.

(v) Finks

Earlier, we saw masks as a potential problem for the simple conditional analysis and for our modal ability approach to understanding. But there is another case which forms a problem for the simple conditional analysis of dispositions, namely *finks*. Martin (1994) says that some dispositions are "finkish" in the sense that the disposition could be acquired or lost within the circumstances that would normally serve as the stimulus. In other words, there is an intrinsic property causally responsible for the disposition, but "this intrinsic property (the causal basis) is lost, after the object suffers the stimulus but before the response comes into being." (Bird, 1998, p. 227) For example, imagine the glass turns to lead just as it is struck because there's an overprotective sorcerer who likes his glasses fragile, but not broken. "A finkishly fragile thing is fragile, sure enough, so long as it is not struck. But if it were struck, it would straight away cease to be fragile, and it would not break" (Lewis, 1997, p. 144). If you want to entertain a more realistic example, consider an electric saw with a sawstop. The saw can cut off fingers, but whenever it would do so, there is a failsafe mechanism that prevents it from doing so.

In like spirit, we may say that a subject's (epistemic) abilities can be finked. These finks can be externally applied (e.g. someone violently distracts the subject by pushing her whenever she is asked a question so she can't respond, or even concentrate) or intrinsic to the subject (e.g. the subject is prone to panic attacks).[98] Nonetheless, finks are finks because they make a difference to what would normally occur. This has lead authors who support a dispositional analysis to try to keep out such finkish counter-examples by either (a) restricting the circumstances to only those where finks are absent (e.g. Choi, 2008), (b) focusing on standard circumstances (also called "ceteris paribus") where they would presumably be absent (Steinberg, 2010), or (c) by pointing to a suitable proportion of circumstances under which the disposition does occur (e.g. Manley & Wasserman, 2007). None of these suggestions is without its criticisms (Choi & Fara, 2018), which may spell bad luck for the hope of providing an exhaustive analysis of the concept of "disposition". Fortunately, we are not trying to provide such an analysis, we are merely trying to figure out whether the subject will respond appropriately under a range of salient circumstances. If the subject acts appropriately under standard

[98] A reverse situation can also be imagined, where the ability isn't finked, but is finkish, meaning the person doesn't have any abilities until she gets tested, and then someone or something makes it so that the subject does respond appropriately. We'll come back to reverse finks in Section 3.3, when we discuss abilities without understanding.

circumstances, then the understanding attribution is uncontroversially warranted. If we can ascertain whether the subject *would* act appropriately under a suitable range of circumstances, that would still validate the understanding attribution. And if we can't, but we can identify finks, we can decide whether circumstances that exclude them are salient to consider (e.g. is it relevant to understanding that no one distracts the subject? Is it relevant that the subject doesn't get panic attacks?) or whether excluding them from circumstances is worth doing (e.g. is it too much to ask people distracting the subject to leave? Is it too much to buy or allow anxiety medication?). Bringing counterfactuals doesn't just draw open the space of possibilities, it also allows us to detect, consider and assess such difference-makers. Crucially, when we determine that subjects with finked abilities still warrant understanding attributions, it is not *in spite* of abilities, but because of it.

Like with masks, the crux of finking cases is in the difficulty of evaluation, and in the limit of (salient) circumstances (i.e. those without the fink), not the absence or irrelevance of abilities under salient circumstances. This is not to say finks are always equally easy to detect[99], but the mere fact that we can clarify what a fink is, shows we have a means of pointing to the difference in abilities *as a fink*, while retaining abilities as a mark.

(vi) Tools

In the quote of Wilkenfeld (2013a) with which we started this chapter, one of the offered examples was that the "inability to do derivations in first-order logic could arise from a broken pencil" (p. 1003) If someone needs a pencil to be able to do derivations, but doesn't have one, do we have a genuine case of a subject with understanding that nonetheless lacks the appropriate (epistemic) ability? Let's take a closer look.

The epistemic ability in question here is that of making derivations in first-order logic. It seems fair to assume that it is implicit that the circumstances are otherwise quite standard (and we may even assume, for the sake of argument, that standard circumstances don't involve pencils). It also seems fair to assume that the subject fails to make derivations in first order logic even if she tries to, which is just a way of saying that the lack of action is not due to her being unaware of what is expected of her or her having no interest in doing derivations. Instead, it arises from a broken pencil. But if it arises from a broken pencil, then is it not also implicit that if the subject were to have a pencil (which is not broken), she would showcase the appropriate abilities by performing the appropriate derivations? If she didn't, why is the broken pencil the difference-maker out of which Wilkenfeld says that the lack

[99] Furthermore, a mental state approach has no easier time at it, since we can only derive mental states from acts.

of abilities "arise[s]" (p. 1003)? Either we are attributing understanding without any indication of why it is warranted (and what the broken pencil has to do with anything) or the warrant for understanding comes from abilities in a range of counterfactual circumstances (where there is a whole pencil), which means that it is once again the abilities marking the understanding.

Furthermore, if we can find a pattern of counterfactual difference-makers (e.g. the lack of an unbroken pencil), then it is perfectly possible for us to contextually devalue (or diminish) the salience of a range that excludes this difference-maker (e.g. it is okay to expect a pencil being present), as opposed to others (e.g. it is not okay to expect a textbook with the proof in it being present). The salience of these difference-makers can vary depending on who is doing the attributing. We need to have a systematic way of deciding which (kind of) differences are salient (and in Chapter 2, I have made conceptual room to do so), but we don't need to look for something behind the acts or abilities. Salience of circumstances or no, when the attribute is deserved, it is abilities that mark the understanding.

But do they really mark the understanding of *that subject?* Fara (2008) considers cases such as that of the broken pencil as one involving masking. But it is not quite as clear here whether the concept of "masking" is indeed appropriate. In the pencil case, it is not the presence of something masking the ability, but the absence of something (e.g. a broken pencil) masking the ability. Since human subjects are not entities that routinely come with pencils attached, it is perhaps unfair to say that the lack of pencil masks the subject's understanding. Furthermore, a case could be made that the target of the attribution was supposed to be the human subject, sans pencil. That case is clearer if the tool were more sophisticated (say, a program which helps making logical derivations - although it preferably can't do it by itself, just like a pencil can't either). In that case, it would be better to say the subject+tool pair has the ability, but the human subject by itself does not. I can deal with unsophisticated tools, such as pencils, by either considering it as a *fair background condition* in similar ways to room temperature (which, under my approach, means we can consider those circumstances that involve a pencil as the most salient range) or *as an extension of the subject* in a similar way as hands are (which we can conceptualise as part of the salient economy-parameter). The use of sophisticated software, on the other hand, does require us to deny understanding to the human individual (because it is not a fair background condition and more than an insignificant extension). For our present purposes, all that matters is that the problem with the use of logical derivation software is not that we would need to demarcate understanding beyond abilities, but that we would need to demarcate the subject beyond the human individual. Why and how is something which I will consider briefly in Section 3.3.iii and more extensively in Chapter 4, where I conceptualise extended understanding.

(vii) Skill Deficiency

The next example of understanding-without-abilities we will consider comes from Kareem Khalifa. In arguing against de Regt's skill condition of understanding, Khalifa (2012) says that we can attribute understanding even in cases where there aren't any skills present, as long as there is the appropriate (explanatory) knowledge. He says:

> "Understanding amounts to (a) knowing that the explanans is true, (b) knowing that the explanandum is true, and (c) for some I, knowing that I is the correct explanatory link between the explanans and the explanandum." (Khalifa, 2012, p. 26)

For example: to understand jet lift based on the Bernoulli principle, the subject must "know Bernoulli's principle and the jet's initial conditions (the explanans) and facts about the jet's lift (the explanandum), [and] she must also know that the explanans entails the explanandum" (p. 26). If we accept that satisfying Khalifa's requirements leads to understanding (which I will), a subject can understand jet lift based on the Bernoulli principle without having the ability to successfully use the principle (which is what de Regt's skill condition required - at least for scientific understanding). This may seem like we therefore have a case of understanding without abilities. But only under a narrow view of the abilities involved in understanding. The disagreement merely comes from focusing on different types or kinds of abilities, such as (a) applying the theory, and (b) explaining the theory. Knowing involves believing, and we have seen that beliefs are only beliefs because they lend predictive or explanatory power to a subject's acts. If we can attribute the belief that the explanans is true and that the explanandum is true, then presumably the subject is able to do quite a few things, such as for instance endorsing both the explanans and explanandum, give justifications for why they are true, explain the used concepts, rephrase or translate the principle, etc. Both the acts that make us attribute "skill" and the acts that make us attribute "belief" (even if there's not much overlap) will be appropriate acts that warrant the understanding attribution. So this does not quite shift the mark of understanding as defended here.

Furthermore, my contextual approach allows me to acknowledge that de Regt's skill condition (requiring the ability to use the theory) may be too high a standard for all contexts of understanding attributions (although it certainly seems appropriate for certain scientific contexts). And thanks to my scope parameter, I can say that the lack of ability to use the theory (in the way the skill condition requires) does not necessarily undermine the understanding attribution, but that its inclusion would further the degree of understanding. Moreover, what warrants the understanding in both cases, can once again be found in (different types of) abilities.

(viii) Bad Luck

As one last counterexample, what if a subject understands but simply has some bad luck. Imagine Watson. They are an expert in their field, and up until now, they have always acted appropriately. But Watson has a bad day, and all of their attempts and acts just happen to be inappropriate (be it through an unlikely miscalculation, an unlucky gush of wind, a misunderstanding or an unlikely priming of the wrong answers due to the order of the questions). Such bad luck does not make us assume that Watson has lost their understanding. Is this a case of understanding without abilities?

It is not hard to recast such claims about bad luck in the wider perspective of modal abilities. Bad luck is bad luck because the lack of the appropriate acts were *unlikely*, given the (rough) circumstances. In other words, under most salient circumstances that are roughly the same, the appropriate acts would not be lacking. So under a wide range of counterfactual circumstances, Watson would display their ability. Furthermore, the difference-maker between the failing in actual circumstances and success in counterfactual circumstances would be slight non-epistemic changes (such as, for example: a different gush of wind, the questions asked in a different order, a little coffee, etc).[100] This marks an important difference with a subject who has simply *lost* their ability, because that subject would no longer display their ability in any of the wide range of salient circumstances (except those that provide epistemic changes, such as cheat-sheets, lots of time to study, etc).

Only if the subject doesn't keep failing (e.g. under similar circumstances) would the understanding attribution be vindicated, and would the (previous) actual displays be dismissed as mere bad luck. But once again, it is abilities that vindicate the understanding. If the circumstances under which the subject fails were more prevalent, however, we wouldn't call it bad luck, but a bad ability, and if the subject fails under nearly all circumstances except one, it is actually *good* luck. Which brings us to the notion of abilities without understanding.

3.2 Abilities without Understanding Objections (Lack of appropriate acts)

So far, I have been dismantling counterexamples that seem to suggest it is possible for there to be understanding without abilities. But if abilities are the true mark of understanding, as I have argued in Chapter 1, then it would also need to be impossible to find a case where we can discern the appropriate abilities, but cannot attribute understanding. To present a genuine counterexample, one is required to motivate that abilities are present, but the understanding attribution is undeserved.

[100] What's more, Watson would, in the subsequent days (allowing a little time, involving no epistemic events such as lots of studying), again display their ability in the wide range of salient circumstances we expect of them.

Motivating such an example won't be easy. We've already seen that we need to be careful with intuitions at first-sight, overly speculative psychology or incoherent metaphysics. And because we saw in Chapter 1 that understanding attributions are regularly warranted through abilities even under accounts that do not place their premium on abilities, we better take those lessons on board. Furthermore, the parameters (discussed in Section 2.1), as well as their salience (discussed in Section 2.2) and evaluative features (discussed in Section 2.3) can now be taken into account to judge the strength of the ability or abilities. Having emphasised the *scope* of abilities, it can become more difficult to attribute understanding on the basis of a single ability. Having emphasised the *sensitivity* of an ability, it can become more difficult to justify there being an ability based on a single type of act. Having emphasised the *range* of acts (expressed through counterfactuals) it can become more difficult to justify there being an ability based on an appropriate act in a single circumstance (or a limited set thereof). Having emphasised the robustness of the acts, it can become more difficult to justify there being an ability based on current circumstances. However, if we do find convincing arguments that what we discern are abilities, but no warranted attribution of understanding, we would need to remove abilities from their throne as the mark of understanding.

This section (as well as the next one) will cover several candidate counterexamples that (seemingly) showcase the presence of ability, but a lack of understanding. The literature on education, philosophy of mind and epistemology have offered up several such candidate examples, and I will consider them (and others) here. I have divided the candidate counterexamples into two sections, because I believe there are two distinct ways in which they fail to counter the ability approach as presented here, namely by (a) failing to show that the appropriate abilities are indeed present (or indeed lacking), and by (b) attributing the understanding or abilities to the wrong subject. The former cases I will present in this section, and the latter cases I will present in the next one. So, in this section I will try to show that the lack of understanding in each one of the proposed counterexamples can be recast as direct or indirect claims about the lack of abilities. In doing so, I will showcase the fruitfulness of being able to recast understanding attributions as the salient scope, stability and sensitivity of abilities.

(i) The Lucky Shot

Firstly, what if someone was just *lucky*? There are multiple ways in which one's appropriate epistemic acts can be due to luck. An act could be a one in a million shot (the lucky shot), could have easily or usually been inappropriate (environmental luck), could be appropriate for a reason that would usually be appropriate but isn't now (Gettier luck), or was unlikely to be acquired at all (evidential luck). Some authors are willing to accept that understanding can be compatible with some types of luck (e.g.

Kvanvig, 2003; Pritchard, 2010). Others (e.g. Khalifa, 2013) mark understanding in such a way that lucky understanding is impossible. Where my account differs most from those presented in the literature is that I won't draw a clear distinction between luck and "non-luck". As such, my ability-based account does not keep luck at bay in the same way that an anti-luck condition is intended to - but given the incessant string of counterexamples to each anti-luck condition, neither do they. What my account does manage to do, which anti-luck conditions don't, is present us with a way to always point to the problem of luck as a lack of the appropriate abilities in some way (depending on the type of luck in question): as a lack in range, scope, robustness and/or the salient sensitivity. This is because they function as degrees in parameters rather than as conditions. This will be shown to be a strength of my account, not a weakness. Furthermore, it entails there's no hard divide between luck and "non-luck," which I will argue is a strength as well.

Let's start with the first type of luck, *the lucky shot* (as I will call it). There are countless examples where a subject acted appropriately (usually when that act is neither routine nor easy), but that act was just a stroke of good luck, a one in a million success. We saw this exemplified in Glick's quote in Chapter 2. Here it is again:

> "Suppose the novice trampolinist's new coach asks [her] which tricks [she] is already able to do. The correct answer would not be a massive list including every trick [she] could pull off given some incredible *stroke of luck*." (Glick 2012, p. 129, italics added)

The easiest epistemic example of a stroke of luck or lucky shot is a student giving a correct answer through a wild stab in the dark.[101] The luck involved can vary with the specificity of the ability tested. If a student of mathematics guesses the correct answer in a multiple-choice exam, she was a bit lucky. But if she randomly filled in "x = 24.36" on a math-problem merely because it felt right, it would be incredibly lucky if it turned out to be correct. If on an astronomy exam, a student constructed a sentence of random letters (like the infamous monkeys on their typewriters), forming: "due to the motion of the light-emitting objects in which objects moving towards us, light is shifted to the blue, and for objects moving away, the light shifts to the red" and was correct, it would be unlikely on the verge of implausibility.

[101] The stab in the dark is metaphorical if we're talking about epistemic abilities, but literal if we're talking about her ability to hit a precise target with a knife.

What immediately becomes clear from these examples, though, is that what makes something luck is how *unlikely* it was for the subject to respond appropriately. The reason why it is lucky is because the causal base of the lucky shot *doesn't* usually result in the appropriate acts (outside of an incredibly narrow set of specific circumstances such as this one). So the act cannot be repeated or would not have occurred if circumstances were different, sometimes even if they were only ever so slightly different. If it did, we wouldn't so readily call it luck. But if it was luck, then the subject's counterfactual or future successes will be sorely lacking. And it is not unreasonable to place the threshold for understanding beyond a few successful acts in very precise circumstances.

The degree of unlikeliness can go up with the degree of accuracy (explaining the doppler effect by a randomly generated string of letters is unlikely) or scope (getting all the answers correct in a multiple-choice exam by guessing is unlikely) tested. But even in cases of lucky scope, scope would only be possible under a very specific set of circumstances. Outside of those circumstances, luck would quickly run out (if it doesn't, it wouldn't be considered luck anymore). Once we draw open the scope and sensitivity of appropriate abilities, the problem of luck as a lack in abilities becomes clear. If someone guessed a number as the solution to a math problem, can they also solve different problems in the same vein? Can they explain one of the steps in the reasoning? Can they make an analogy with another problem? Cite the relevant theories or formulas? Find mistakes in an answer? Presumably not. If they can, why call it luck? Similarly, if someone strung together letters to define the Relativistic Doppler Effect, can they also give examples? Can they explain it visually? Can they relate it to other theories? If they can, why call it luck? Any method that consistently gives appropriate results would be hard to call mere luck. But where is the cutting off point? It now becomes clearer why the dividing line between luck and non-luck can be difficult to keep as a hard edge. Luck is a matter of degree, not kind.

In the case of a lucky shot, the circumstances of the test are misleadingly positive and the appropriate acts wouldn't have been present in nearly any other circumstance (or not enough of them anyway). The lucky shot is the opposite of our earlier example of bad luck. They both provide misleadingly unlikely circumstances to test the ability. The difference is that in the case of bad luck, the circumstances fail to show the full scope, sensitivity and/or stability of the ability present and in the case of the lucky shot, the circumstances fail to show the *lack* of scope, sensitivity and/or stability of the ability. So the diagnosis of this situation is not a problem with the narrow success, but with the wide failing it entails. The lucky stroke is lucky only because abilities are lacking. As such, it is not a counterexample where we find abilities without understanding. Nevertheless, the lucky shot is not the only type of luck.

(ii) Environmental & Evidential Luck

Environmental luck is in many ways similar to the lucky shot, except that it adds the little snag that the kind of stability, scope or sensitivity tested for is not usually salient, but the environment makes it so. Consider the *Barn Facade example*, as explained by Pritchard (2009):

> "Here we have an agent who sees a barn in clear daylight and so forth and, using her reliable cognitive abilities, forms a belief that what she sees is a barn. Moreover, this belief is true and is not gettierized since she really is looking at a barn (...). Nevertheless, her true belief is epistemically lucky – in the sense that she could have easily been wrong – because unbeknownst to her she is in barn facade county where nearly all the barn-shaped objects are in fact fake barns which are indistinguishable to the naked eye from the real thing." (Pritchard, 2009, p. 26)

When it comes to the Barn Facade example, it is my contention that the two most important features are a *lack of sensitivity*, and the *contextual salience of that sensitivity*. Let's begin by pointing out something about the latter feature. The (contextual) salience of the sensitivity in this example is unusual. Most contexts of attribution interested in "understanding what a real barn looks like" will involve a sensitivity which involves distinguishing non-barns (such as sheds, houses or castles) from barns, as opposed to barn facades from full barns. In that sense, the environment (being in a barn facade county) shifts the context of attribution, because a new sensitivity (distinguishing barn facades from real barns) becomes salient. So it is crucial to point out (which I have seen no one do) that this change in environment implies a change in context of attribution.

Next, let's consider the lack of sensitivity. The lack of sensitivity is revealed even in most authors' explanation of the example. Pritchard (2009) says "she could have easily been wrong" (p. 26), Grimm (2006) points to "how easily your belief might have been false" (p. 519), Khalifa (2013) adds that "had Bonnie looked at a fake barn, she would have believed it was a real barn" (p. 3). In short: the problem is not that she cannot recognise a barn, it is that she lacks the ability to tell a fake barn from a real one. She is not sensitive to the difference (or her scope does not include distinguishing the two). If she happens to be correct in recognising a real barn in barn facade county, then the circumstances under which you tested her ability were misleadingly positive and not generalisable to other counterfactuals (i.e. where she is faced with a barn facade). In that sense, once the contextual salience is included, the situation is not inherently different from the lucky shot.

There is another type of epistemic luck that seems to be a case of environmental luck, when the luck stems not from the luck in the ability itself, but from its acquisition. Imagine Nero came to believe[102] that her house burned down due to faulty wiring because that's what a firefighter told her. If it is true, she would gain the ability to relay a correct explanation. So far so good. But what if, unbeknownst to her, some people nearby happen to be dressed up as firefighters for a costume party. If she spoke to a *real* firefighter, she was lucky, *environmentally* lucky, because she acquired the true belief (i.e. its related appropriate acts) from a reliable source even though she might equally not have. This is the case of *environmentally lucky Nero*.

Pritchard (2009) argues that environmentally lucky Nero does understand even though she was lucky. I am inclined to say that Nero, under this particular version of the example, doesn't understand (even if she does know), but that this is not because of her environment. Nero doesn't understand even if she talks to a real firefighter because all she can do is relay "it is because of faulty wiring." Following a firefighter's advice may give you knowledge of faulty wiring causing a fire, but it doesn't give you understanding of it (in exactly the same way as the memorisation case - see subsection iv). Understanding should be composed of a larger scope than that. But this is easily fixed. Let us say that the firefighter actually explained it in great detail. She explained how the house was wired and where the faulty wiring was, she taught her how wiring works, how this faulty wiring led to an issue, what could have been done to prevent it, etc. Here I can follow with the intuition that Nero understands, even if she was lucky to.

Yet it is intuitive to still find something unnerving about coming to understand if one was so lucky. Those intuitions become more pressing if at any given point, there were 100 fake firefighters for every real one. However, if we would try to keep out environmental luck of this kind because it was lucky the appropriate belief was *acquired*, we would also keep out the epistemic luck of any unlikely evidence (i.e. evidential luck), which is more obviously benign. In fact, Khalifa (2013) criticizes Prichard by saying the Nero example "doesn't involve environmental luck; only evidential luck." (Khalifa, 2013, p. 18). To substantiate, he brings up the example (which also appears in Pritchard, 2005) of being able to glimpse who robbed the bank. Just because a successful glimpse was unlikely doesn't entail that it hurts the understanding or knowledge acquired by it. My adapted Nero case, who acquires a broad scope of abilities from the real firefighter, makes it even clearer why her luck is *evidential luck* and therefore relatively benign.

[102] As a reminder to the reader: I will be operating with an interpretationist approach to belief (see Section 1.4), which entails that believing something (e.g. that the house burned down due to faulty wiring) implies certain acts from which we can derive that belief (e.g. relaying that the fire was due to faulty wiring when asked why the house burned down).

But we haven't fully addressed what's unnerving about such luck. I therefore present the following analysis: In ordinary circumstances, following the advice of anyone with the signifiers of an expert, such as being dressed as a firefighter, is a relatively sound way to guard yourself from gaining inappropriate abilities or losing appropriate ones. But in a sea of fake firefighters, that becomes naive and, even if it leads to the appropriate abilities, the abilities will be anything but robust. So what is lacking here is not the subject's sensitivity (the fake firefighter can be as extensive in her teachings as she is correct), but the *robustness* of Nero's abilities. If Nero follows the advice of any person in an authoritative coat, she will lose her abilities just as quickly as the danger of running into them is high. In other words, the abilities will not be very robust. Understanding should be made of sterner stuff. What we need to guard ourselves against such unnerving cases is a robustness parameter for the subject's understanding, not an etiology condition. If our subject is too easily swayed, she will be equally easily swayed to give up the right abilities by the wrong source of information, and its is the ease of swaying that is the problem for understanding (even if the wrong source of information is the problem for knowledge).

Of course, this problem only becomes salient if the risk of it is high enough. Under most circumstances (where there is no sea of convincing fakes), being swayed by anyone in an authoritative coat will lead to the appropriate abilities being relatively robust (because only experts will sway). But in a sea of convincing fakes, the lack of robustness becomes particularly salient.

Furthermore, this robustness can be linked to other abilities. Here, Khalifa's (2013) insistence on being able to evaluate explanations (which we can subsume as one of the contents in an understanding's scope) becomes particularly salient, because not only is this an ability that can be tested in its own right, it is also an ability that will improve the robustness of the appropriate abilities. To substantiate, Khalifa presents two further subjects: one which can identify experts and one which can evaluate explanations. Both understand better than Nero did, but "when it comes to understanding, explanatory evaluation is a more important ability than identifying experts." (p. 13) This makes sense, because identifying experts leads to a higher robustness of the appropriate abilities (because only convincing fakes would be believed), whereas explanatory evaluation leads to a higher robustness still (because good explanations of experts without signifiers would still be believed, whereas even convincing fakes wouldn't be).

Once again, it is abilities which truly mark the understanding, and not the source or mode of acquisition - and if they do, it is only to the extent that those play a role in the presence or lack of

abilities (be it in scope, sensitivity or robustness). But what if the fakes are luckily providing us with appropriate abilities? This brings us to Gettier luck.

(iii) Gettier Luck

In the field of epistemology, a Gettier case is a type of counterexample to the justified true belief account of knowledge. It relies on situations where the subject has a justified true belief but can nonetheless *not* be ascribed with knowledge, thereby showing that the three classic conditions for knowledge are insufficient. It is relevant to us here because a similar case could conceivably be made for understanding: if the subject in question doesn't warrant an understanding attribution, but does seem (luckily) to justifiably believe something true (and therefore has at least a set of salient abilities), then we have a case of abilities without understanding. I will argue, however, that *Gettier luck,* to the extent that it is a valid problem, is still an ability-centred problem.

Gettier examples originated with two cases provided by Edmund Gettier (1963). Consider Gettier's first case: Suppose that Smith and Jones have applied for a job interview. Smith has strong evidence that Jones will get the job and that Jones has ten coins in her pocket – for example because the president of the company told her she is very fond of Jones. From this, she can infer that "The person who will get the job has ten coins in her pocket." She believes this and she believes it justifiably. Now imagine that Smith is actually the one who gets the job. Unbeknownst to her, she also has ten coins in her pocket. This means that Smith's justified belief that "The person who will get the job has ten coins in her pocket" is still true. And yet, so Getier (1963) argued, we can't say she knows it.

While I do believe knowledge can be gettiered, I don't think Gettier's cases are Gettier cases.[103] Gettier's case actually becomes more ambiguous when seen through the lens of the intentional stance mentioned earlier.[104] Here is why: According to the intentional stance, if Smith believes that the person with 10 coins in her pockets will get the job, then this is not because there is a proposition "The person with 10 coins in her pockets will get the job" which Smith is supposed to stand in a relation to, but because there is a belief-ascription of "the person with 10 coins in her pockets will get the job" which is supposed to be explanatory or predictive of Smith's behaviour. But, as the case shows, that ascription is explanatorily misleading and predictively incorrect. The ascription of the belief "The person with 10 coins in her pockets will get the job" will be predictive or explanatory in certain ways, but wrong in many others. If Smith finds out that she had 10 coins in her pockets, she doesn't now

[103] I have not found anyone making similar arguments, so I will tread carefully here.

[104] Due to the dominant focus on beliefs as propositional attitudes, the intentional stance based argument I will present is one I have not come across in the literature, even though it has strong ramifications.

behave as if she will get the job. If she gets the job, she will not now assume she has 10 coins in her pockets. If anything, the belief-ascription functions almost exactly like the ascription of the belief "Jones will get the job.'' The latter ascription is therefore more explanatory and predictive, and therefore more warranted. While it is true that Smith may infer new beliefs from old beliefs (for example, beliefs related to who possesses 10 coins and how that stands in some relation to the person getting the job), those beliefs are not solitary propositions, but ascriptions of the subject that are part of a larger web of ascriptions which together have explanatory and/or predictive power. So the belief "The person with 10 coins in her pockets will get the job" is at best ambiguous and at worst incorrect. The same argument can be made for Gettier's second case, although I won't make it here.

Conceptualising the quality of understanding with the scope, stability and sensitivity parameter makes the problem all the clearer. While it is true that Smith has some abilities (e.g. she can endorse the true prediction that the person with 10 coins will get the job), the lack of stability, sensitivity and scope of abilities will uncover her low degree of understanding. Firstly, the abilities supplied by Gettier cases are not robust. As soon as she finds out she has 10 coins in her pocket, she will disavow her prediction that "the person with 10 coins in her pocket will get the job" because it doesn't distinguish between her and Jones. So her understanding is not appropriately stable. Conversely, if she found out Jones had no coins in her pocket, she would adjust her prediction accordingly. So her understanding is not appropriately sensitive to the situation. Lastly, her understanding does not have an appropriate scope. Even if we leave the threshold relatively low (don't require of Smith that she has the ability to explain how the candidate got the job, why that candidate was better than the other, when it would have been otherwise, what the process of contacting the selected candidate involved, what is expected of the candidate once hired, etc), there are still a cluster of relatively trivial abilities that we expect from someone who understands "who got the job". So we cannot say that Smith understands who got the job if she describes the selected candidate as "Jones" instead of "me," explains the reason why the candidate got the job as her "being favoured by the president" (which is true of Jones, but not her), and doesn't show up to work because she didn't know that was expected of her. The degree and dimensions of the appropriate abilities involved in understanding even a relatively simple situation such as this one quickly can quickly reveal the lack in understanding.

But when people talk of understanding being gettiered, they usually use other types of examples, and not just those suggested by Gettier himself. Pritchard (2009) talks of Gettier cases as those where "something intervenes 'betwixt belief and fact'." (p. 21). The ability is appropriate (and even

appropriately acquired), but is nonetheless lucky because what would usually make it appropriate is ill-suited to the situation. Here is a classic example:

> "Suppose that our agent is looking into a field and, using her reliable cognitive abilities, forms the belief that there is a sheep in the field. Suppose further that this belief is true, but that the agent is not in fact looking at a sheep but a big hairy dog which looks just like a sheep, and which is obscuring from view the sheep that is in the field. The agent in this case clearly lacks knowledge since it is just a matter of luck that her belief is true. Nevertheless, she is forming a true belief via the stable and reliable cognitive abilities that make up her cognitive character." (Pritchard, 2009, p. 21)

A common way to deal with such luck is to require anti-luck conditions for knowledge.[105] But while knowledge may require an anti-Gettier condition because it deals with singular beliefs, understanding is inherently broader. Knowledge seems to suffer from keeping out the problems of luck exactly because it lacks the degrees of scope, sensitivity, stability that make understanding so resistant. As such, the problem of luck is acutely felt for knowledge and it makes more sense to require its justification to be very origin-bound. However, understanding, as I have presented it, can bracket its origins and deal with the problem through its dimensions and degrees. For instance, Kvanvig (2003) argues that:

> "understanding requires, and knowledge does not, an internal grasping or appreciation of how various elements in a body of information are related to each other in terms of explanatory, logical, probabilistic, and other kinds of relations that coherentists have thought constitutive of justification." (Kvanvig 2003, p. 192-193)

Kvanvig subscribes to a misleading metaphorical language which leads to the wrong mark of understanding (see Chapter 1), but the gist of his argument is not far removed from my own: namely that understanding requires a scope, stability and sensitivity that will uncover the problem of Gettier cases without requiring an additional condition. The reason that a Gettier example does not undermine the ability-account, is because the *lack of understanding* in Gettier cases is due to a *lack of abilities*. Merely endorsing that there is a sheep in the field will not suffice even for understanding something as simple as that there is a sheep in a field. For such understanding, we would also find it

[105] Pritchard (2009) avoids an anti-luck condition by instead considering the subject through virtue epistemology. Here, the subject's success needs to be *because* of her reliable abilities.

appropriate that the subject also be able to tell us the rough location of the sheep, its physical appearance, why it is visible to her, etc. Even though the subject is lucky, her luck does not extend beyond the appropriate count of 1 sheep. In other words, her scope of abilities is severely lacking. Moreover, if it is lucky that she was able to give the right count of sheep in the field, she will fail in a significant amount of nearby counterfactuals (e.g. in every counterfactual where the dog does not obstruct the sheep, she will count two sheep, not one dog and one sheep). The range of appropriate abilities will be limited to the circumstances where the dog shields the sheep from view. Furthermore, the salience of distinguishing dogs from sheep now becomes especially clear (in much the same way that it did in environmental luck cases). And even though we are assured by Pritchard of her usual "reliable cognitive abilities," this unusually salient sensitivity is not included. Clearly, her reliable cognitive abilities are not strong enough in range, scope or sensitivity. Therefore, the example does not undermine the ability account of understanding, even if it does undermine the traditional account of knowledge. Understanding is harder to fake, because it is made of broader stuff.

So what if knowledge were made of broader stuff, and thus closer to understanding? Consider Knowledge-how. Knowledge how, even if it reduces to knowledge-that, is broader than knowledge-that. Furthermore, knowing-how seems to distinguish itself from knowledge-that exactly because it is less susceptible to Gettier problems, as was pointed out by Stanley & Williamson (2001) in their original article on knowledge-how. And yet, Stanley & Williamson don't claim that knowledge-how is immune from Gettier cases. In fact, they still offer an example for how it can be gettiered:

> "there are indeed Gettier cases for knowledge-how. Bob wants to learn how to fly in a flight simulator. He is instructed by Henry. Unknown to Bob, Henry is a malicious imposter who has inserted a randomizing device in the simulator's controls and intends to give all kinds of incorrect advice. Fortunately, by sheer chance the randomizing device causes exactly the same results in the simulator as would have occurred without it, and by incompetence Henry gives exactly the same advice as a proper instructor would have done. Bob passes the course with flying colors. He has still not flown a real plane. Bob has a justified true belief about how to fly. But there is a good sense in which he does not know how to fly." (Stanley & Williamson, 2001, p. 435)

The diagnosis that Bob does *not* know how to fly is not readily accepted, however. See (Poston, 2009) for a convincing dissenting view. But the relevant question for this dissertation is not whether Bob *knows* how to fly a plane (and therefore whether know-how can be gettiered), but whether Bob

understands how to fly a plane (and therefore whether understanding can be gettiered). The Bob example is a good one because it already presupposes the scope and sensitivity of abilities that were absent in the Nero or sheep example. Does Bob not understand how to fly a plane? The question can be formulated differently: If Bob is put in a real plane and flies it successfully, does he, through the experience, *gain* all of his understanding of how to fly a plane, or does the experience merely vindicate his understanding? Intuitively, it seems to be the latter. At worst, one can say that Bob's understanding has increased a little because he is now able to justify the applicability of his flying technique, but the technique itself was already understood. Standing by this intuition keeps understanding as a coherent and useful concept based in abilities that is furthermore distinct from knowledge-that (and possibly knowledge-how).

Now let us consider the *acquisition* of the ability being gettiered. Now we come back to the Nero case. Imagine *Nero II* came to believe her house burned down due to faulty wiring when a firefighter told her, and it was true. She would again gain the ability to relay that correct explanation. But unbeknownst to her, she was talking to a fake firefighter in fancy (but convincing) dress. Bad luck. But there's also good luck, because (as a wild guess to upkeep the illusion of their fancy dress) they happen to give her the same short "it is faulty wiring" answer that, luckily, the real firefighter would also have given. Nero I was environmentally lucky enough to talk to the real firefighter and Nero II was Gettieredly lucky in talking to the fake one, but as a consequence both are equally able. In Nero II's case, she is gettieredly lucky because she acquires the appropriate belief (i.e. its related appropriate acts) from an unreliable source and therefore could have easily been inappropriate.

Pritchard (2009), who has also turned this example into a Gettiered version, argues that gettieredly lucky Nero II doesn't understand because of the unreliability of the source. What is central to this way of keeping out luck is the focus on how the knowledge, understanding or abilities were formed or *acquired.* I am generally very averse to marking understanding through how it was acquired, because it will lead to discrediting abilities, no matter how robust or sensitive, unless or until they are validated by what are considered the appropriate channels - and how do we determine those if not by their success in granting appropriate abilities? This may be a fruitful avenue to pursue in the characterisation of knowledge, where abilities are less directly relevant to its justification, but understanding's heavy reliance on abilities to justify the understanding (even outside of ability-accounts) makes this avenue less fruitful. But how do we keep out such unjustified origins if all we can point to is the presence or lack of appropriate abilities?

The reader may already anticipate why I don't think Nero II would understand - the same reason the first Nero didn't: a severe lack of scope. If all that Nero can do is repeat what the firefighter told her, then even if that ability is appropriate, she is no better than the case of memorisation (to be discussed next). But we can easily fix that again by creating a stronger ("scopier") version of the Nero II example. Imagine if the fake firefighter gave an elaborate response, explaining the wiring of the house, where the faulty wire was, how faulty wiring can lead to fires, why it happened in this case, how the problem could have been detected earlier, which things would have made a difference, etc. All of these explanations were wild guesses about Nero II's situation, but they happen to be correct. It would (perhaps) be fair to say that the beliefs Nero II acquires from it are not knowledge.[106] But does Nero II lack understanding in the same way she lacks knowledge? Once she acquires the same information from a trustworthy source, her true beliefs would turn to knowledge, but what happens with her understanding? Does it makes sense for her to look back on her younger self, lamenting her lack of understanding contrasted with her recently acquired understanding, even though she navigated the issue with the same scope, sensitivity, range, etc as she does now? If it changes anything, I would say the uncovering of a justified origin-story validates her understanding as applicable to her situation, but it doesn't make the understanding *appear*.

Pritchard (2009) also addresses a case that is like the stronger ("scopier") version of Nero II, namely Kvanvig's (2003) case of the *book on the Comanche nation*: Imagine that someone is trying to understand the history of the Comanche nation through a particular scholarly book, and all the information in that book is accurate, but it was based in complete guess-work on the part of the author. This is similar to the stronger version of Nero II. Nevertheless, even here, Pritchard believes this would lead to a lack of understanding. Why? Because of the similarity with Nero II (the weak version) talking to a fake firefighter. But that comparison is unfair, because the person learning about the Comanche nation through the book was luckily granted a much greater degree of scope than the weak Nero II case where the fake firefighter merely got the cause correct. Therefore, it is not clear why the intuitions of one case should so readily traverse to the other. At what point would the subject start to understand? For instance, does she start understanding each point of information once she doesn't get corrected by an expert in the room? What if the book gets taught at a university by someone who has done extensive research on it? Does this validate the book as a reliable source for understanding? Does it do so for every copy in the world or only those in that professor's class?[107]

[106] I'm not giving an account of knowledge, nor does my account of understanding depend on one, so I'm leaving it open whether this is true or not.

[107] Even for knowledge, Brogaard (2005) has the opposite intuition to Pritchard: "Arguably, the anti-luck requirement is a bit overblown. If you learned quantum mechanics from an unreliable textbook, you might still have mastered the

(iv) Memorisation

A case which was both raised in Wilkenfeld's quote (at the beginning of the chapter), and which I have mentioned a few times so far myself, was that of rote *memorisation*. It is quite easy for a student to study a few of the answers to common questions verbatim and then relay these answers when the questions sound similar to the ones they learned the answer to. Most people will readily agree to the claim that memorisation is not the same as understanding, and yet it is hard to deny that the student has some ability, namely the ability to answer those questions. It may seem that such a case forces us into one of two avenues: deny the subject has any ability, or shift our position where the subject's understanding is marked by the appropriate abilities. But this is a false dichotomy. We do not need to discredit the abilities that are present, nor do we need to credit the student with understanding based on this ability. So what is going on then?

Firstly, it is worth noting that the student who memorised the questions-and-answers does have more abilities than the student who can't answer any questions at all because she has never even interacted with the material. If forced to compare the understanding of both students, the former would rate higher than the latter. Nevertheless, neither of them have enough abilities to warrant an understanding attribution. There is a threshold of understanding that simply has not been reached. For every question-answer pair that the student has studied, there are countless answers to countless questions that the student is unable to give. Even if we assume that the student doesn't fail as soon as the question is phrased slightly differently (which is a limitation on her range), the number of questions that the student will be unable to answer, or problems that the student will be unable to solve, vastly outweigh those that she can (which is a limitation of sensitivity and scope).

If the student happens to get exactly the same 10 questions as the 10 question-answer pairs she studied, then the problem is not abilities without understanding, it is good luck. We can see this as soon as we draw open the possibilities of testing by considering counterfactual tests that involve different questions or, if you don't want to stray quite so far from the factual, by adding a lot more questions. Not dissimilar to the lucky shot, the circumstances under which the subject can answer correctly are limited to those where the questions asked overlap with the question-answer pair that were memorised. So the ability does not extend much beyond a certain set of stimulus-response pairs. If the student got lucky with the questions, then that is a problem of narrow evaluation for wide abilities (and called a kludge - see Section 2.3), not a problem of the mark of understanding lying

theory, be a world-class expert, be able to answer any questions on the topic, and so on. So, it seems that you can still count as knowing [and therefore understanding] quantum mechanics by every standard that matters." (p. 17)

beyond the abilities. The student's failures to act appropriately are directly related to the student's failure to understand. On the other hand, the student's successes are directly related to the student's understanding (although it will still be too narrow to reach the thresholds for most contexts of attribution). It is not the narrow *success* of rote memorisation, but the wide *failing* it entails that makes for a poor understanding.[108]

(v) False Beliefs, False Theories & Idealisations

Another type of counterexample, also raised by Wilkenfeld (2017), is to be found in cases where the subject displays competence on the basis of absurd or false beliefs. For instance, imagine that a subject can predict the trajectories of planets, but thinks that all of the planet's motions fundamentally reduce to a theory about the planets being pushed and pulled by angels. To a much lesser extent, something similar can be said about competences on the basis of idealisations. For instance, imagine that a subject has some abilities related to certain economic trends, but her modelling is based on individuals being rational utility maximisers. The basis of these subject's abilities aren't just subtly false, they are absurdly false. They may have an adequate scope of stable abilities that are sensitive to variations, but the understanding attribution is on shaky grounds.

The strength of such counterexamples ties into the question of whether understanding requires a *factivity* or *veridicality* condition. Does the subject (and her theories) need to be strictly "true" or appropriate to understand? de Regt & Gijsbers (2016), who also address this counterexample, believe understanding can be achieved independently of the veridicality of the scientific theories used to understand. The difference between Newton's mechanics and an angelic theory of planetary motion, for instance, is not that the former is a representational device which is (approximately) true while the latter is false. (In fact, even Newton mechanics isn't strictly true either). Instead, the difference is that the former is (more) scientifically effective - meaning it is intelligible and leads to reliable success. Its use tends to "produce useful scientific outcomes such as correct predictions, successful practical applications and fruitful ideas for further research" (p. 1) Such an effectiveness condition also makes it clear that the appropriate theories for understanding will vary, both in a pragmatic (it depends on whether the scientist can use the theory) and contextual way (it depends on whether it advances the current research practices). This does justice to the contextual nature of understanding, which a veridicality condition would have more trouble dealing with. A focus on effectiveness, however, would entail that Ptolemy understood planetary motion, even though he endorsed a geocentric model. But

[108] In Section 6.1, I'll consider the extent to which regressability (to the author of the answers to be memorised) does entail that the attribution may not *uniquely* apply to the student.

it is a feature, not a bug, that a contextual approach allows us to attribute (some) understanding to past scientists who used theories even though they have proven false in some way. The veridicality condition by contrast, would deny understanding to anyone (no matter how effective their theory) up until the moment a true theory comes along.[109] Furthermore, due to the high reliance on idealisations or idealised models in the sciences, we must either accept idealization as part of scientific understanding, or give up a lot of understanding attributions. (Baumberger, 2011) I largely agree with de Regt and Gijsberg, but will try to give some extra weight to the point by showing that even a veridicality condition is not wholly distinct from claims about abilities.

Firstly, I must note that part of the problem of factivity comes from being focused on propositions or representations. If one conceives of understanding as propositions stored in the subject, then it seems natural to link the quality of understanding to the correctness of the stored propositions, as Wilkenfeld (2017) does.[110] This gives plausibility to the idea that one could perform adequately while operating under only an approximate or even downright incorrect propositions (or other representations).[111] Then, we can discern appropriate abilities, but deny understanding on the basis of the incorrect propositions behind them. But in Chapter 1, we have argued against the proposition characterisation, except as an instrumental concept. Under the ability account, a belief must be discerned through the subject's acts and abilities, so the distinction between understanding and not understanding cannot be a private mismatch between a correct and an incorrect proposition. Instead, the distinction must reduce to inappropriate or insufficient abilities. Therefore, the problem with false beliefs isn't the appropriate abilities associated with them, but the failures they entail. For instance: If Hilde has a reasonably good understanding of most causes of greenhouse gasses, but believes livestock grows on trees and therefore have an impact similar to (or lower than) the production of plants, such as corn or soy, then her abilities related to the climate crisis (explanations, predictions, teachings, and policies) will suffer as well. Her predictions of the amount of greenhouse-gasses will be severely skewed and her proposed solutions will speed up the climate crisis rather than temper it.

The case of Hilde was relatively straightforward, because she had a false belief which directly affected a large set of the appropriate acts we are looking for. But what if false beliefs were less consequential?

[109] We have also moved on since Copernicus and "[w]e readily agree that Copernicus did not know the Earth's orbit to be circular, but it seems inappropriate to entirely deny him understanding of the planets' motion. It is also quite a stretch to say that idealized models do not provide understanding, strictly speaking." (Baumberger et al, 2016, p. 7)

[110] Wilkenfeld (2017) phrases it as "representational-accuracy" (p. 1273), where the accuracy of the representation can be seen as composed of, but also broader than, true beliefs. What exactly representational-accuracy is, is not explained, but "the general idea is that the actual state of affairs of the world is in some important sense similar to the state of the world as depicted in the representation." (p. 1275)

[111] Indeed, Wilkenfeld (2017) prides himself on it being an "ability-free view" (p. 1274)

For example, imagine that Titiana can predict planetary motion (including what would happen if the planets changed course), but believes that the motion of the planets is creditable to invisible fairies pulling them around. This is (presumably) false, but, unlike Hilde, her false belief doesn't affect much of her predictive or problem-solving abilities. Does this mean that we finally have a case of abilities without understanding? No, because the lack of understanding is still equal to the lack of appropriate acts. It is true that if Titania were tested with an exam on planetary motion filled with what would have been different questions, she would respond appropriately. But if the exam asked how gravity works, she would respond with a literal fairy-tale. Furthermore, if the exam asked (for example) what would happen if there wasn't an ether, she may respond that the planets would stop moving because the fairies can't navigate without it. Or if she were asked about what happens in a vacuum, she might respond that the planets stop moving because the fairies can't breathe. Both of these responses are not only inappropriate, they come at the expense of lacking the actual appropriate responses.

Once again, the problem of false beliefs can be restated as claims about a lack of appropriate abilities. It is true that some of these abilities would only get detected in very narrow circumstances (perhaps the fairies would never come up, except in highly specific metaphysics questions), but to the extent that they would or should, it is in those circumstances that the problem lies. One may wish to insist that those highly precise metaphysical questions make all the difference - but if they do, then surely they'd be a salient part of the understanding's scope, and of the examination.[112]

But can't an entire theory be strictly and absurdly false, while still supplying appropriate abilities? For instance: Let us say the object of understanding is human psychology. Our subject, Agnes, can predict that the insecure army general with an ambitious partner will become morally reckless, that the old merchant who faces social oppression will look for economic-based revenge and that the young princess who has just lost a parent will be prone to self-pitying soliloquies. Agnes can make rough predictions of individual behaviour which are significantly better than wild guesses, so she clearly has some abilities related to human psychology. So far so good. But if you were to press Agnes on the underlying causes, it would be revealed that she relies in large part on the concepts of humourism and the four temperaments: a psychological theory based on the balance of four chemical systems (black bile, yellow bile, blood and phlegm) regulating human behaviour. According to Agnes, an excess of

[112] Of course, the appropriateness of metaphysics may be hard to justify. If there were no way to distinguish the gravity-based theory from the fairy-based theory then science also can't favour one theory over the other. But if it can, there are also inappropriate acts for which points can be subtracted. Furthermore, the smaller the effects that the "false belief" has on the set of salient abilities, the harder it is to detect as a false belief - either because it is unclear why that belief is false (due to its lack of making an empirical difference), or why the subject believes it (due to its lack of effect on the subject's behaviour).

black bile is to be found in the princess, too much yellow bile is stored in the merchant and blood flows too freely in the general. But the psychological theory of humourism and the four temperaments has long been relegated to proto- or pseudo-science. On the basis of this, we might be inclined to deny our subject with understanding. Nonetheless, it is worth noting that the reason the theory is considered to be false will correlate with the reason the subject lacks abilities. She may prescribe the army general with blood-letting, the old merchant to stay out of the sun and advise the young princess to avoid cold meals - none of which will make a significant difference. Additionally, she may have rough estimates of the ratio in how much black bile, yellow bile or blood we would find in each of them (if we were to cut them open) - and she would be proven wrong upon empirical inspection. Outside of her accurate predictions, she has a whole scope of abilities which are lacking, inappropriate, insensitive or inaccurate. If they weren't, we would have to reconsider our position. If her predictions of behaviour and descriptions of anatomy, as well as her prescriptions of how to regulate them were always spot on, we would have to reconsider humourism (or at least a close correlate) as a scientific theory for the same reason that we would also have to reconsider Agnes as an understanding subject. The problem with Agnes is not with the humours-based abilities she has, but with the wide failing they entail. de Regt & Gijsberg (2016) make a similar point about understanding on the basis of hepatoscopy (predicting the future of plantary motions on the basis of livers):

> "If it is indeed by pure accident that the livers and the planets are always aligned, then the theory of hepatoscopy is not reliable (...). Hepatoscopy therefore gives no understanding, which means that our intuitions and the effectiveness condition are in step. If, on the other hand, the livers and the planets are robustly aligned, in what we might call a lawful way, then hepatoscopy would be reliable and effective, and would give understanding." (de Regt & Gijsbers, 2016, p. 61)

In other words, if livers and the planets consistently aligned, it would become less obvious why hepatoscopy would be unscientific.

Idealisations, however, fall somewhere in between. If the idealisation is purported to deviate from "the truth" (i.e. something more appropriate) in some way (thereby making it an idealisation), this deviation must translate into inappropriate acts (e.g. incorrect answers, inaccurate inferences, misplaced skills), because if it doesn't, how would we know it deviates from "the truth"? If a subject uses an idealisation in the appropriate circumstances, she will succeed in the same way that the idealisation succeeds. But if she uses it in situations where the approximation is inappropriate, she

will fail in the same way that the idealisation fails.[113] However, if she stops short of making these mistakes as soon as the idealisation meets its limits[114] (e.g. she stops using Newtonian Mechanics involving relativistic speeds), she will understand more (because she has the added ability to point out where the idealisation fails). And if she, in those areas, swaps her idealisation for a more sophisticated theory (e.g. Einstein's relativity theory) or another more appropriate idealisation, then the lack of complete appropriateness is of the idealisations, not the subject. Furthermore, if we allow beliefs to be contextual properties (for someone's beliefs to vary with the situation or circumstances), we could go so far as saying that she doesn't have any false or limited beliefs, because these beliefs are only beliefs when they are appropriate, and in situations that they are not, they are not believed. Crucially, however, false beliefs or idealisation do not undermine the ability-account, because they lack in granting abilities in the same way they lack in understanding. The only problem is, once again, one of evaluation: that we don't always detect the glaring false beliefs, in spite of their salience.

(vi) Short-termed abilities

What if someone has abilities one day, but loses them the next? For example, imagine someone is a quick study, but has a bad memory. A lot of the things that she reads or that are explained to her, she quickly comes to understand and she can display many of the various appropriate abilities we would expect of her in a wide range of (counter)factual circumstances. But the next day, these abilities will have gone. This can be assessed quite simply: She acquires understanding, but quickly loses it - to the same degree that she learned the ability and lost it.[115] As a general assessment, we could say her understanding is not very robust, because she'll quickly lose it. And it is only in robustness that she's failing, nothing else.

(vii) Employed Algorithms or Models

Next, consider a case where someone uses a *rule, algorithm or model* to successfully perform without even knowing or contemplating what they are doing. Skemp (1976) called this phenomenon, rather

[113] It may also be that the ways in which an idealisation fails lies outside of the salient accuracy for a particular practice, a particular context of attribution, in which case there is no salient difference between an "idealisation" and "a more accurate theory" to warrant the distinction.

[114] There are multiple ways in which the limits of the idealisation can be respected: not endorsing the idealisation in every circumstance (range), for every question (width), if a better approximation would be offered (robustness) or is necessary (sensitivity).

[115] One may be tempted to think that what we have here is the opposite of temporary impairment: temporary competence. If that were true, then should we make the opposite assessment? Namely that the subject didn't understand before or after, and therefore not during either, even in the presence of abilities. But it would be an odd assessment to say that the subject's lack of understanding is masked by learning these abilities. The concept of mask is not usually intended to measure someone's lack of understanding. If we valued lack of understanding more and attempted to lower people's understanding, then the mask narrative would make more sense. So perhaps the situation is not quite symmetrical because our interests regarding understanding aren't either.

succinctly, "rules without reason" (p. 20). But if we mark understanding by its abilities, should this absence of reason (in the use of rules) worry us? If not, why not? All of us have, at least at one point, come at an appropriate answer by using a rule, while having no idea how or why it worked. For instance:

> "[I]t seems that someone might very well have the propositional knowledge that f = ma (again, imagine [s]he comes to this knowledge via reliable testimony) without understanding or grasping, as it were, how the law ''works'' – a point that, as Philip Kitcher observes, will be all-too familiar to science teachers who have seen students do well on the rote portion of a test while nonetheless doing poorly on the application portion, where they are supposed to apply their grasp of the law to particular cases (see Kitcher 1989: 437–8)." (Grimm, 2012, p. 7)

This quote about Kitcher captures the situation of the example pretty well. But it has also already pinpointed the problem. Even while the diagnosis is conceptualised as a lack of grasping, the evidence is a lack of abilities. The student who uses "rules without reason" will struggle to apply it to examples, to draw the link with other problems, to draw quick inferences about strange exceptions, to point to limits of the algorithm, to adapt the formula for different intentions, etc. If the use of rules, algorithms or models only leads to a lack of understanding, we would leave them out of education and science altogether. See (Kuorikoski & Ylikoski, 2015) for a defense of how using models can increase understanding.[116] The difference between someone who understands and someone who doesn't isn't whether or not a rule was involved or used, but how it was involved or used. In short, it is a problem of scope and sensitivity.

But there is more to be said about this objection, which will be made clear when you consider the following pushback: What if the algorithm used is complex enough to broaden the scope and sensitivity sufficiently? We have good reason to believe the user doesn't necessarily understand anything she is doing, even though she is able to display a large amount of understanding in the answers produced with the algorithm. The problem here is no longer the lack of scope, but the incorrect targeting of the appropriate subject. We will come back to this when we address "blind rule following" in Section 3.3, but the gist of it is that the algorithm and subject come apart. Interpretationism can help us conceptualise that there are answers that belong to the subject, and

[116] In Section 3.3 and Chapter 4, I will explain why the use of external representations can force us to extend the subject to include the external resources. Part of this point can also be found in (Kuorikoski, 2011).

answers that belong to the algorithms, because the two can be disentangled from one another too readily to attribute understanding to the subject (rather than the algorithm, merely implemented in the subject).

(viii) Abilities from Emulation

What if we could think of entities that match us in abilities, but where the entity itself makes us wary to attribute understanding? Van Camp (2014) cites a popular counterexample just like this:

> "It could be argued that understanding always goes hand in hand with ability, but that does not demonstrate that they are identical. We may frequently judge that a person understands according to their abilities, but we can also judge a computer or emulating robot by the same standards without concluding that it understands." (Van Camp, 2014, p. 98-99)

The counterexample here claims that (i) computers or robots can *emulate* the same abilities, but that (ii) this won't be enough to warrant an understanding attribution. I believe the intuitions behind this counterexample benefit from misleading assumptions on both parts of the claim.

Consider the first part: "computers or robots can emulate the same abilities." Van Camp (2014) makes this claim in the context of epistemology, but it is a popular stance in other fields as well.[117] What is intended with "emulation"? Is it a complete replication which adopts all or most of the entity's aspects or behaviour, or does it only replicate certain aspects which are assumed to be relevant? If only a narrow set of the aspects that are assumed relevant are replicated, the limits of the system can become clear as soon as it ventures even slightly outside of its narrow area of intended competence. A calculator does very well on complex multiplication, but it doesn't take much to figure out its abilities do not stretch very far beyond it. One minute a computer surprises us with its cleverness and creativity[118], but a moment later, it may do something so nonsensical or superficial, that we feel that we were duped by a cheap trick (or at least only a moderately priced one). I called this a kludge (see Section 2.3). In calculators, both the type of ability (e.g. rote calculation) and amount of abilities (e.g. not much beyond it) misrepresent what we usually look for to attribute understanding most (i.e. the

[117] The most notable example is Searle's (1980/1985) critique of what he calls "The Brain Simulator Reply" (p. 363)

[118] This applies better to simple automated theorem provers than to calculators. Consider the pons asinorum proof found by Gelernter's program, which showed that the angles of an isosceles triangle are equal by noting that triangle ABC is congruent to triangle ACB (i.e., its mirror image) (Hofstadter, 1999). While it can certainly be called a creative move, it is not as "out of the box" a method for the program as it would have been for a human. So the level of creativity we infer from it is misleading.

content and thresholds of the parameters we find most salient). To the extent that this is the argument, the denial is fair, but the assessment is incorrect. What stops us from attributing understanding to computers is not the fault of it being a computer, but due to its failures of competence. The problem with the calculator, for instance, is not that a calculator doesn't really calculate (whereas human mathematicians do), it is that it does little else. Its capacities aren't fake, but our assessment based on them can be mistaken or misleading. So when there is a lack of abilities, the computer doesn't warrant the understanding attribution. But not because it is a computer or robot, but because it lacks the appropriate abilities.

Importantly, my argument for why these computers don't understanding (i.e. because they lack the appropriate abilities) doesn't rest on the claim that computers or robots *inherently cannot* replicate the appropriate abilities (I don't believe we have evidence for this claim – see Chapter 6), but on the claim that they usually don't, and if they do, only narrowly (and misleadingly) so. This does not entail that they never will. But it may misguide our intuition when we are sceptical to attribute computers or robots with understanding. Once again, what is guiding our intuitions is the lack of scope. Nevertheless, that is only part of the problem.

What if computers or robots are more sophisticated than they are at present and could perfectly emulate the same abilities? Now we have satisfied the first part of Van Camp's claims and come to the second part, namely that "we can also judge a computer or emulating robot by the same standards without concluding that it understands." Let us say that an expert is emulated in a robot down to every last detail. This entails that the emulation will have every possible ability that the human expert has. Personally, I am very willing to give it every attribution of understanding that the human expert has, but other people feel there is something iffy about that. If there is, it is up to them to clarify, beyond the iffy feeling, what the relevant difference is. We will get more into this line of argument in later chapters, but suffice it to say that, even if there are great differences (e.g. the hardware is silicone instead of organic), it is difficult to argue why *that difference* is a relevant one for withholding understanding attributions. It may seem intuitive that computers or robots do not warrant understanding, but is the intuition strong or sound enough to be effective without further clarification? If we should come to a different conclusion with two entities judged by the same standards, then I believe a stronger argument is required for why there should be a double standard.

One candidate argument for the double standard is that the computer or robot only has *derived* abilities, because the abilities really belong to the programmer and not the robot/computer/program.

But is this a fair assessment that contrasts where humans derive their abilities from and how? If computers or robots derive their abilities, why do human abilities not derive from the guiding hands of teachers and/or to the process of natural selection? In Chapter 6, where I address artificial understanding, we will take a closer look at what it means for an ability to *belong* to a particular entity, but suffice it to say, it is hard to argue for a decisive difference between humankind and machinekind that is also relevant to understanding attributions.

3.3 Abilities without Understanding Objections (Wrong Subject)

We have now come to the second type of counterexamples that involve claims of abilities without understanding. I will argue that each of these objections (save the last two) involves a failure to target the appropriate subject that possesses the ability. This avenue of defence, as well as the conceptualisations that go along with it, is largely lacking from the literature on understanding (and to a less dramatic extent from the literature on knowledge). Therefore, I will give a rough outline of the situation here, and give the subject its due attention in the last 3 chapters, where we will expand on each of the counterarguments to the objections raised here in more detail. To end, I will also discuss a case where it is mistakenly presumed that the understanding is attributed to the wrong subject, and a case where it is mistakenly presumed that the lack of coherence needs to be conceptualised and addressed in the subject with understanding, rather than through the object of understanding.

(i) Mimickers & Marionettes

Consider Echo. Whenever you ask Echo about tomorrow's weather, she will always give you an accurate prediction, as well as further explanations on how to make these predictions and where to find the appropriate data. She will even offer corrections on where you made mistakes as well as offer little relevant titbits about meteorology, generally. In short, whichever scope of abilities you are after, she will supply it. What's more, she will provide this under most circumstances you'd ever meet her in. Abilities are clearly present. And yet I am here to say she does not understand anything about meteorology. Why not? Because what most people don't know is that Echo is married to Ororo, and Ororo is an expert meteorologist who is in constant communication with her wife, Echo. They are connected via bluetooth at all times, and whatever question Echo gets, Ororo answers through this bluetooth so Echo can repeat it to the person asking.

Does this mean we have a case of abilities without understanding? Only if we allow a *change of subject* between the attribution of abilities and the (lack of) attribution of understanding. While it is quite uncontroversial to say that Echo doesn't understand and that abilities are present, are the abilities

those of Echo or those of Ororo? It is quite easy to find out. Consider all those counterfactual worlds where Echo is not in contact with Ororo and you will soon find that there is a severe lack of abilities in Echo. Ororo, on the other hand, can answer all of the questions posed to her even without being in contact with Echo. Clearly, only one of them is doing the heavy epistemic lifting. If we would want to read the situation as abilities without understanding, it is Ororo who has abilities, but Echo who lacks understanding. Echo does display the same abilities, but only by echoing Ororo. When we are targeting a subject to attribute with abilities or understanding, we are making implicit assumptions about where the abilities are implemented. And just about all of the relevant implementation for the abilities is here outsourced to Ororo.

We can take the situation even further. Consider the following situation, based on Peacocke's (1983) puppet or Dennett's (2009) marionette: Suppose we found a subject that answers every meteorology question appropriately. But if we were to surgically open that subject, we would find only radio transceivers. The lifeless body is simply controlled as a radio-controlled puppet by something or someone off-stage. If it is controlled by a brain or machine that doesn't control any other bodies, then there is no need to "change the subject." It is merely a subject whose brain is kept in a non-traditional location (akin to the thought-experiment in Dennett, 1978b). If, on the other hand, some evil or neutral scientist controls this body along with that of her own, then the abilities and understanding we attribute really belongs to the scientist instead.

The complex multi-track abilities of an understander have to be realised somehow and somewhere. But wherever it is realised, that is where the abilities and understanding can be attributed to. We will delve deeper into the subject with understanding and how it is realised in Chapter 4, but what is crucial for our current purposes is that the problem with these types of cases is not that a particular subject can have abilities and lack understanding, but that between the attribution of abilities and that of (the lack of) understanding, the target-subject was changed.

(ii) Reverse Finks

Now consider a related situation: The subject doesn't have any abilities, but whenever the subject is tested, someone else makes sure that she has the causal base to respond appropriately. Essentially, this is the *reverse* case of *fink*, it is a positive fink, one that makes one "gain" abilities. A classic example of this is an evil scientist who has some means of control over another subject – for example, a device planted in her brain that controls her body. Whenever the subject, let's call her Shaw, is about to be tested, then, and only then, the evil scientist activates the device and makes her react appropriately.

As soon as the act is performed, the device is turned off again and Shaw is left to her own devices.[119] Does the subject understand?

If the scientist is constantly monitoring Shaw and implementing the appropriate reactions manually, then the scientist must have the appropriate abilities so as to explicitly translate them into device-based operations that make sure the subject acts in the appropriate way. In this case, the role of the scientist is actually that of the puppet-master, like in the marionette case. This would result in a misevaluation of understanding, but only because we would be targeting the wrong subject. As soon (and as long) as we incapacitate the scientist or its device, Shaw will act with the same lack of abilities as we deny her quality of understanding. In demarcating the subject who the abilities belong to, the difference-maker is the scientist and nothing else, so the implementer of the abilities is unambiguously the scientist.

But what if the device were more autonomous? What if the device was once designed by an evil scientist, but is now outside of the scientist's control and a permanent part of Shaw. Would Shaw understand? Now, the question of whether she understands is more of a question of whether the mechanism is "external" to her or not. If the line that separates her from the external world is drawn between her brain and the device, she (sans device) will continue to lack the ability and thus the understanding (although the device won't - more on that in Chapter 6). But if the device is sufficiently part of her to be "internal", she understands because the device is now part of "she". There actually are good arguments for considering the device as internal, because the circumstances of the device springing into action are no different from the appropriate brain-mechanisms doing so. More on that in Chapter 4, where we will explore the notion of "external resources" and its "external" modifier in more detail, as well as develop how to demarcate a subject. But what is crucial for our present purposes is that the counterexample again only seems to work if we target the wrong subject to attribute the (lack of) understanding to.

(iii) External Resources

A next candidate counterexample is one where abilities are displayed by the subject, but only when that subject can make use of the relevant *external resources*. For instance: If someone uses a

[119] The set-up of this counterexample is very like those of Frankfurt (1969), dubbed *Frankfurt-style cases* in the debate on free will. Those discussions focus on whether the subject "could have acted otherwise" (e.g. see Fischer, 2002), whereas we focus on whether the subject acts appropriate and who is responsible for those acts. If the reader is worried that Frankfurt-style cases are a problem for free will (and therefore understanding) I would like to redirect them to my response to them in the appendix of (Delarivière, 2015).

barometer to predict the weather, does she understand storms? Wilkenfeld (2013b) certainly doesn't think so:

> "[E]ven if I owned the best barometer in the city and was only interested in storms so that I could predict them, it still sounds odd to say that I thereby understand storms." (Wilkenfeld, 2013b, p. 92)

He makes sure that there's no issue of scope misguiding our intuitions by reducing the salient abilities of understanding storms to their prediction only. So, according to Wilkenfeld, even if we limit our understanding attributions to just one relevant ability, namely that of predicting storms, we would still be averse to attributing understanding.

de Regt & Dieks (2005) also address the barometer counterexample, and do so under their ability-account. They appeal to the scientific-theory requirement in their criterion for understanding phenomena (CUP). Either reading barometers does not involve a scientific theory, and therefore doesn't satisfy their CUP-criterion for scientific understanding, or, if it does involve a scientific theory (one which embeds the correlations of barometers and air pressure), it is no longer clear why their use couldn't lead to understanding. We can translate this into a general (in)ability-claim as well: One the one hand, if the only ability our subject can display is a (fairly inaccurate) prediction of storms on the basis of reading the text (e.g. "rain") the barometer points to, then the subject is not only lacking in scope, but also outsourcing all of the relevant work to the barometer. If our only interest was in predicting storms (which of course, it isn't), I believe Wilkenfeld should attribute (the minimal amount of) understanding to the barometer. The prediction is one of the barometer, not the subject. This means that the subject is, in this case, no more than a mimicker (addressed in subsection i). Except that predicting storms is usually too narrow a scope for understanding attributions, which is why we don't usually attribute barometers with understanding. On the other hand, if the subject has a scientific theory, then this implies that the subject does have a wide scope of abilities, which would reduce the role of the barometer to mere data or input about air pressure and no meteorologist is held responsible for sensing air-pressure by themselves. This would be allowed as a background condition (i.e. considered as a salient resource under the economy-parameter).

Perhaps barometers are too easy to dismiss because with the meteorologist their contribution is too meagre to count as more than input, and with the amateur their contribution is too narrow to change the subject from the amateur to the barometer. We have seen that cognitive outsourcing can be so

extreme that everything that is relevant can be outsourced (e.g. the mimicker) and we can agree that there are forms of cognitive outsourcing which are trivial or benign (e.g. barometers), but what happens in cases where the cognitive load is more evenly divided?

Let's consider a case where the external resource plays a larger role. If Mathilda has a smartphone with meteorology-tool apps at her disposal, she could conceivably predict the weather and adequately answer many questions even if she wouldn't be able to without the app.[120] Some of the causal base of the ability is outsourced to the app, but only some of it. The ability is present, but, unlike in the case of the mimicker or the marionette, I cannot simply change the subject from Mathilda to the smartphone (even if it was a subject or agent), because the smartphone cannot predict the weather or answer any questions by itself any more than Mathilda can. And not just because it needs Mathilda to copy in the data, but because it wouldn't even remotely know where to look for the relevant data, or how to convert it to the appropriate form for the app to do its work, etc. The ability only comes when Mathilda and the smartphone "work together." This may be a case of an extended understander (a topic which we'll come back to in the Section 4.4). The subject that can be attributed with abilities, is not the human subject, but the pairing of the human subject with the external resource.

Of course, even in the Mathilda case, we can contextually devalue the importance of cognitive offloading or outsourcing, given the economy weights within the economy parameter. If pen and paper are relatively easy to come by, we might not care so much whether the subject with understanding is Mathilda+pen+paper, because we can allow or expect Mathilda, if she is required to perform, to have pen and paper readily available. Similarly, when meteorologists get recruited, their abilities are not tested as a pairing between them and a particular smartphone, because the use of such an app (or a similar one) may be granted anyway. They become "background conditions." What is still relevant here though, is that these background conditions do some of the cognitive lifting. Therefore, the "subject with understanding" is realised in Mathilda and her app.

(iv) Giant Look-up Table

In the memorisation case, I left open the possibility of an extreme version where every possible question-answer was accounted for. Imagine every possible question-answer pair were stored in a book or a program. Because the book or program contains the appropriate response to every question (and every follow-up), the person who uses it or has memorised its content will be able to respond appropriately to every question. In essence, it is a *giant look-up table* combined with something or

[120] A similar case is presented in Ylikoski (2014).

someone to look up the appropriate response in each circumstance. (Dennett, 2009) The example is similar to the robot sketched by Block (1981), but is even closer to the *Chinese Room thought-experiment*[121] as conceived by Searle (1980/1985). Because the book is as extensive as it is, it can supply a broad scope of appropriate answers with the necessary sensitivity to variations. So there is a wide scope and sensitivity of abilities (every input has an appropriate response), but is there understanding? If it is, who is the subject? If it isn't, what is wrong with the ability-attribution?

One glaring problem with this counterexample is, of course, that it is physically impossible. This "imagined system would be a computer memory larger than the visible universe, operating faster than the speed of light. If we are allowed to postulate miraculous (physics-defying) properties to things, it is no wonder we can generate counterintuitive 'possibilities.'" (Dennett, 2009, p. 347)[122] Furthermore, because it is impossible, it can only remind us of the one thing that can plausibly come close to it, and that is the memorisation case, which did have a lack in scope. So we should be wary of relying solely on our real world intuitions in assessing an impossible counterexample. Nevertheless, I will continue to address it.

Let us start from the assumption that there is, in actual fact, a wide scope of abilities, and therefore understanding. Whose is it? Because the counterexample often focuses on the *user* of the giant look-up table, it is usually assumed that that user is the only appropriate target. Is there any other possible target? Perhaps the book - if the book is used, then we may find the case similar to the mimicker case. But even if the book is doing most of the epistemic lifting, it is still incomplete as a subject. It needs something or someone that looks up and responds what the book tells it to. It needs a CPU.[123] The subject is the physical or virtual system *as a whole*. This analysis is essentially the *System Reply* or

[121] Searle (1980/1985) objects to a premise of the Turing test, namely that a performance model of consciousness (and understanding or interpretation) is sufficient to mark the property. To make his point, he proposes the following thought experiment: Suppose Searle is in a room receiving pieces of paper from the outside world with Chinese sentences on them. He does not speak Chinese, so to him they are no more than meaningless squiggles. However, Searle has a very large instruction-book (written in English, which he does understand) to produce the appropriate squiggles in response. Upon completing the procedure, he slides the response back into the world. If the instruction-book contains sufficiently sophisticated procedures, anyone outside of the room (who does understand Chinese) will get the impression that an Chinese-speaking person is responding from inside the room. Pry though they might, the responses will be absolutely indistinguishable from those of a native Chinese speaker. However, regardless of the success of these responses, it would be ludicrous to say Searle, by virtue of the instructions, understands Chinese because he couldn't interpret the symbols. (Searle, 1980/1985; Heylighen, 2014)

[122] A similar point can be made about Oracles, who just *know* all the answers without any scientific theories or cognitive information processing. This makes them a similar case of abilities without understanding. But "[o]ur hypothetical oracle is just a figment of the imagination: in reality one can only make successful predictions if one understands the relevant theories." (de Regt, 2017, p. 107)

[123] "What's the fun of life if we're not being processed?" says Achilles in (Dennett & Hofstadter, 1985, p. 447). The dialogue it features in is incredibly relevant to the Giant Look Up Table objection, because it centers around the possibility of having a conversation with Einstein if all we have is a book detailing the process of Einstein's brain.

Virtual Mind Reply to Searle's Chinese Room thought experiment (Searle, 1980/1985). The system reply is analogical to what I have already been arguing for so far (namely that the appropriate subject is the whole implementing system), and the virtual mind reply is analogical to one which we will consider in the next objection (subsection v), so I will deal with it then (and expand on it in Chapter 4).

So the Giant Look-up Table counterexample is one that is physically just not possible, meaning our intuitions may be contaminated with what is physically possible, namely the Memorisation case. And if the counterexample were physically possible, it is clear why we can't target the book as the understander, but it is not clear why we can't target the book along with an implementer as its CPU (which is another form of the external resource objection).

(v) Blind Rule Following

When I addressed the employed algorithm objection in Section 3.2, I left some stones unturned. I will now turn them by considering a case that is a bit like an internalised giant look-up table. It is not hard to imagine a situation where a student, let us call her Henrietta, memorises a set of rules of a theory, and then uses those to successfully answer questions even though she doesn't really understand what she is doing. If we are right in asserting that "Henrietta doesn't really understand what she is doing," which is an uncontroversial claim, the ability-account would need to address why the displayed abilities do not constitute understanding for her.

Such concerns might be behind de Regt's (& Dieks, 2005; 2017) emphasis on qualitative skills over exact calculation. By disallowing the reliance on exact calculation, it is true that he would keep at bay most of the problems from this counterexample. But, in contrast to de Regt, I believe disallowing exact calculation or algorithmic procedures excludes the wrong feature. To motivate this, I will instead focus on subject demarcation and the scope or sensitivity parameters to keep the problem of this counterexample out. So Henrietta either doesn't have the appropriate abilities (much like the memorisation problem) and/or isn't the one who the abilities belong to (much like in the external resources problem, except the resources are not physically external, but virtually external). Let us address each in turn.

Firstly, we may find that the rule-system is insufficiently complex to do justice to the scope of understanding that we would expect to find in a true understander. When Henrietta has memorised a few formulas verbatim, she might be able to solve some of the textbook problems that require straightforwardly applying that formula (giving her a big edge over the student who has merely

memorised question-answer pairs), but that is also where her abilities will end. She might not recognise cases where they apply outside of textbook examples (which are tailored to the application of the formula). She might not be able to deal with situations that deviate even slightly from the used formula, no matter how slight the change is (e.g. needing to use the formula twice instead of once) or how easy the change is (e.g. knowing whether the absence of information on a variable would require her to fill it in as zero or to leave it out as irrelevant). Recall Skemp's (1976) anecdote in Section 2.3 about the boy who learnt to multiply by dropping the decimal point. Fittingly, I believe Skemp's last sentence in the quote signaled the problem perfectly:

> "He got ten questions right this way (his teacher believed in plenty of practise), and went on to use the same method for finding the exterior angles. So he got the next five answers wrong." (Skemp, 1976, p. 23)

Because the boy was blindly using the rule, the boy suffers from a severe lack of scope and sensitivity in his abilities. So the problem here is not the rule-following, but the lack of abilities it leads to.

But a scope-argument alone does not suffice to fully address all possible cases of rule-following. This becomes clear when we make the rule-system much more complex. So complex that it covers any possible ability one may wish from our subject. If our subject has successfully memorised this rule-system, she will be able to answer any question appropriately. Much like what was the case in the Giant Look Up Table objection, there is no physically possible subject that would be capable of memorising such an elaborate rule-system. But even if it did, we would have to conclude that it is not the student that is the subject that possesses the ability. What we have here is an external resource objection, but where "external" does not mean physically external, but virtually external. There is a simple way to demonstrate the distinction by comparing it to the internalised variant of the Chinese Room thought experiment. In the classic version, Searle, who was inside the Chinese Room, used the book as an external (physical) resource to respond in Chinese. It is quite clear in this case that Searle is reduced to a mere CPU of the book's abilities. But what if Searle learns the book by heart? Then Searle is both CPU and book. Now, we can no longer physically demarcate the subject who understands Chinese as Searle + the book, because *physically*, it is all Searle. But we can *virtually* demarcate (and distinguish) Searle from the Chinese speaker. Who we call "Searle" is not a physically demarcated entity, but a virtual one. Consider asking Searle "what is your stance on the protests in Hong Kong?" If you ask the question in Chinese, Searle will go through his usual motions (except in his memorised rule-book) and give an answer formed by the (memorised) book. Is this *Searle's* stance?

Not really. If you would ask Searle in English, you would likely get a different answer. This is because there are really two virtual subsystems, two pieces of software that happen to share the same hardware (i.e. Searle's body).[124] The virtual Chinese mind is software riding on Searle's virtual (English) mind and they are not sufficiently entangled to unify the two as the same Searle. I will give a more extensive account of what unifies a subject in Chapter 4, but for now, I hope to impart the gist of the distinction in an intuitively clear way.

This is obviously an extreme example, where there are two subsystems that are fully formed selves (with a Searle personality and a Chinese Person personality) that are clearly and unequivocally distinct (which we can distinguish by the full profile of their responses). While extreme, it does, however, help to shed light on the case of Henrietta, who memorised a vast array of complex rules but who does not really know what she is doing. When we consider the abilities, we target an entity which possesses those abilities. In this case, we were inclined to say: it must be Henrietta. Physically, this demarcation makes sense, because the abilities are part of the physical entity we call Henrietta. But the virtual subject is an instrumental entity with beliefs and aims, and those of Henrietta may be distinct from those of the rules she is following (even if only for a time). The beliefs we can discern in Henrietta and the beliefs we can extrapolate from the system of rules (which we can call System Hyde) do not inform one another, so they are not the same virtual system. The problem, therefore, with the Blind Rule Following example is not that the abilities must be discredited, but that the abilities only tangentially belong to the virtual subject we were targeting. The algorithm or theory of rules is like a little blackbox which is shielded from the person blindly using it. This isn't to say that everyone who uses algorithms or a theory of rules doesn't understand what they are doing (the black box can mesh with the brain box, if you will), but there are definitely cases where the answers of the algorithm are distinctly those of the algorithm uninformed by the subject's beliefs, aims or epistemic tactics. The easiest way to show this is when the answers of the algorithm (e.g. the opinion on the Hong Kong protest Searle gives in Chinese) are inconsistent with those of the subject (in related contexts) when she doesn't use that algorithm (e.g. the political opinions Searle gives on Hong Kong in English).

Here's another example: Let us first start with a version of an algorithm that is acting against the subject's desires and beliefs. Let's say that a CEO of a large company has as his intention to maximise profits, no matter what the costs are for other people. To do so, he was given an algorithm that helps him do just that. The algorithm includes all sorts of variables, such as the worker's cost, efficiency and

[124] The situation is slightly more complicated than two virtual systems on one hardware. Actually, it's a virtual system (the Chinese mind) implemented by another virtual system (Searle's mind) – like software running on software.

level of discontentment (assuming each of these could be captured by a variable). This last variable in particular may have a relevant threshold for the interests of the CEO. If the worker is deeply unhappy, they will stop showing up to work, which hurts his profits, but anything above that level of discontentment will no longer have any effect on the profits, and by extension, the CEO's interests. The algorithm may work well enough for most purposes, but unless the CEO understands what he is doing (has further abilities connected to the algorithm), he will fail in reaching his desires in ways that may surprise him. Now let us say the CEO knows nothing about programming or algorithms except that they can be effective. He commissions a programmer to write him a profit-maximising algorithm. But unbeknownst to the CEO, the algorithm he received was actually designed by a benevolent programmer who made sure that the CEO would always share his profits equally with all his workers. The CEO may regularly use and endorse the algorithm, even though it makes him value the level of contentment in his workers far beyond what is necessary for his profits, and even though it makes him pays out his workers with decent wages far beyond the exploitation he would be able and willing to get away with. The mismatch can run deeper still. The algorithm can include assessments of a situation (e.g. we need more workers) along with the appropriate actions (e.g. hiring more workers) which do not align with the CEO's assessment (e.g. "we need fewer workers"), who fires as many workers as possible in person, but then keeps hiring (and rehiring) workers while running the algorithm on his computer. The CEO doesn't understand what he is doing when he uses the algorithm, because the beliefs we instrumentally ascribe to the algorithm (e.g. "Paulina's not been given enough paternity leave to keep contentment high") and the beliefs we could describe to the CEO (e.g. "I am not willing to give paternity leave") don't align at all. Even though the CEO may run the algorithm each day, do we attribute its successes or abilities to him? Or is he merely a tool, a CPU, for the algorithm's abilities? In this situation it seems clear it is more the latter than the former. But here the situation is clear because there is a mismatch in every possible way.

Now imagine the same situation, but with an algorithm, written by a neoliberal programmer, that does exactly what the CEO wants it to do. He is able to maximise his profits (by exploiting his workers as much as is cost-efficient). Is it the CEO who is maximising profits, or the algorithm? They still seem to come apart, because the CEO is at the mercy of the algorithm and not the other way around. If the CEO actually understood the workings of the algorithm (which he doesn't, be it because it is too complex or even just because he never really checked it attentively), he would adjust it wherever it was necessary, correct data where an innocent error was made, recognise when the algorithm ran contrary to his desires (especially if these desires were to change), alter the data fed to the algorithm to fit his beliefs, etc. Here, the algorithm is at his mercy, not the other way around. In instrumental

language, we may say that the algorithm is not *connected* with his other beliefs. This is not a "connection" in the mental realm (e.g. two private entities linked via a Platonic cord). No, it is a metaphorical connection that is justified by the abilities of the CEO to use the algorithm appropriately. In Chapter 4, I will expand on the nature of this metaphorical connection.

Another example: to form a proof in propositional logic, following a system of rules, or even a heuristic, too closely will sometimes result in steps that lead either nowhere (which is a problem of scope) or nowhere where the subject wanted to go (which puts the subject at the mercy of the algorithm's "aims"). The more experienced mathematician would skip it automatically, but the amateur is blindly executing the algorithm or heuristic. The question is not whether that experienced mathematician, at her core, did something *other* than following rules (that is a claim about implementation that is difficult to determine), but whether she was at the mercy of the rule or the rules at the mercy of her. Just because a rule was adhered to does not inherently discredit the success won from it. If someone can successfully perform simple calculations without understanding its symbols or purpose[125], we can say the person is not capable of functional interaction with her own calculation, but we wouldn't say there was no real calculation. The results were simply not *her* results. Likewise, if someone can display appropriate abilities while blindly following rules, we can say that that person is not capable of functionally interacting with the rules she follows, but we wouldn't say there was no real understanding. It just wasn't *her* understanding.

In the case of Searle, the distinction between Searle and the Chinese personality had a clear, hard border, but in examples like that of students blindly following formulas, the distinction between the rules and the student can be one with varying degrees, depending on how much the rules are part of the student's thinking or just running on it. Dennett & Hofstadter (1985) make the distinction more vivid: To say someone understands Chinese involves more than translating the sentences in your head, it involves "mixing the new language right in with the medium in which thought takes place." (p. 379). The reason why the Chinese room stood out so much was because there was absolutely no mixing, no interaction between its homunculus and the system of rules it is following.[126] It is the extreme version of blind rule following, and it makes the crux of the problem all the clearer: the problem was not one of abilities without understanding, but one of targeting the wrong (virtual) subject (see Chapter 4).

[125] The subject might think she's just playing a game according to its rules. There's an interesting example of this in (Hofstadter, 1999). Hofstadter presents a game of his invention, TNT, of which you can learn the rules without realizing it is an implementation of Peano arithmetic.

[126] I hope it is clear that the memorisation case from earlier can fall prey to this exact problem as well. Memorised answers to a predetermined set of questions can betray a different perspective to the rest of the subject.

Most cases of rule-following seem to rely on things like exact calculation. So a good way of testing against blind rule following might be to look for abilities beyond exact calculation – for example, ask the subject to motivate their steps, explain why certain situations will result in errors for the rule or algorithm, and make some qualitative (instead of exact) estimations. If they barely ever (or never) line up with any approximation, there is a clear mismatch. This may be why de Regt places so much power on the ability to predict qualitative consequences, and why he says that "understanding is based on skills and judgments of scientists and cannot be captured by objective algorithmic procedures." (de Regt, 2009, p. 587) De Regt appeals to Brown (1988), who says "explicit following of rules is characteristic of an unskilled, rather than of skilled, performance" (Brown, 1988, quoted in de Regt, 2017, p. 27), which can easily be agreed with. But this should not entail that we need to exclude the abilities that can be captured by objective algorithmic procedures. The question is what the difference is between a skilled and unskilled performer. Is it really the use of rules itself? According to de Regt it is. But does that tactic accomplish what we want it to accomplish?

Imagine if we had two subjects: Zoë and Zelda. Zoë can make very precise predictions based on a scientific theory, and Zelda's predictions are only approximate. In any other respect, they are exactly alike. They can motivate their steps, explain the limits of the scientific theory, correct a misuse of the theory and even adjust the theory and its rules depending on the situation it needs to be used for. The only difference is that Zoë is always precise and Zelda is always approximate. If de Regt is right to insist on qualitative consequences, then Zelda has more understanding than Zoë, because there's no reliance on *exact* calculation. But surely the problem is not that certain abilities (like exact calculation) need to be discredited - for why would exactness be a bad thing? The problem of rule-following, when it is indeed a problem, should instead be found with a poverty of *scope* in abilities (i.e. no abilities beyond the exact calculation) and/or with targeting the wrong subject (i.e. the abilities can be grouped and interpreted as those of the virtual algorithm separate and distinct from those of the virtual subject). Interestingly, to support his position against exact calculation, de Regt says that people can be skilful "without being able to state what they are doing when they do it" (p. 28) and that skills cannot be fully captured or "exhaustively translated into explicit rules" (de Regt, 2017, p. 28). This is very possible, but what this shows is that there are abilities which would be missing from following explicit rules, not that the abilities granted by those rules are the wrong sort of ability or even that skilled scientists never really follow any rules when they arrive at the right answer (after all, they must get at answers somehow). I believe de Regt makes all the relevant arguments, namely urging for a wide variety of skills, for the relevance of practice over blind application of instructions, but I believe he draws the wrong conclusion. The problem is not that using an algorithm or following a rule is

inherently a bad kind of ability, but that some uses of algorithms are done so blindly, that it entails the abilities are limited in scope and/or don't inform the subject's other abilities.

(vii) Derived Abilities

One could object that a giant look-up table is, by its inherent nature, misleading, because it relies on the understanding of whoever or whatever formed the book - therefore its abilities are *derived*. We'll come back to this type of objection in Chapter 6 when we discuss artificial understanders, but I would already like to offer a rough rebuttal. The gist of my rebuttal is that autonomy (i.e. whether the subject can hold up on her own) matters more than etiology (i.e. where the causal line of the subject's acts regresses to). We don't discredit a human subject's abilities as if it is not *their* ability because we can trace them back, fully or in part, to evolutionary processes. Similarly, we don't discredit a human subject's abilities as if it is not *their* ability because they were taught by another subject - not unless they've only learnt a single ability that they blindly implement (in which case there is a lack of scope) or just mimic their teacher (in which case they continuously rely on the teacher, like a mimicker). We also don't discredit a human's appropriate answers to questions because they figured out the answers before they were asked (e.g. by practicing by themselves what they should say when certain questions come up so they can then resort to automatically answering - a kind of self-teaching), not unless they have not retained anything but a memory of specific question-answer pairs (in which case they have now been reduced to a subject with memorised answers with all its scope problems). But if they did retain all of the appropriate abilities to deal with unexpected questions or follow-ups, then the mere fact that she relies on automatic responses for expected questions should not discredit these abilities.

(viii) Lack of Coherence

The last candidate example to counter my ability approach is a bit of an odd one out: Imagine a subject displays abilities, but they are wildly inconsistent. Even though inconsistency is far from what we would expect of someone with understanding, some of the abilities will be appropriate, so could this be a case of abilities without understanding?

One way of keeping out incoherent understanding is by demanding a coherence or consistency condition on the subject's understanding (see e.g. Kvanvig, 2003; Ylikoski, 2009). While I am sympathetic to this idea, I believe such a condition is too strong of a requirement. Even experts can be inconsistent, yet this does not make us withdraw our understanding attributions. Furthermore, the lack of coherence already entails a lack of understanding as incoherence in abilities will automatically lead to some of them being inappropriate. The exception would be if the object of understanding is

itself deemed to be incoherent (e.g. quantum mechanics famously conflicts with the theory of relativity), in which case the coherence requirement would disallow understanding of that object altogether. The importance of coherence should be determined by the context of attribution, not by the mark of understanding. Lastly, a coherence requirement is also superfluous given how my account will approach the demarcation of subjects, which I will expound on in the next chapter.

In Sum

If abilities are the true mark of understanding, as I have argued in Chapter 1, then a substantiated example that showcases we can have understanding without abilities or abilities without understanding would have undermined the ability approach and my account. In this Chapter, I have considered a series of such candidate counterexamples, and shown why each of them fails to hurt the ability approach, as presented in this dissertation.

I first addressed those candidate counterexamples that seem to warrant an understanding attribution, but where abilities seem to be lacking. These involved cases where abilities (i) are masked, (ii) lie outside of non-standard circumstances, (iii) are deliberately avoided, (iv) are (temporarily) impaired, (v) are finked, or (vi) would require tools, (vii) are lacking due to low technical skills or (vii) bad luck. I showed that the presence (or absence) of understanding in each of these cases could be recast as direct or indirect claims about the (counterfactual) scope, sensitivity and stability of the (salient) abilities - thereby keeping abilities in their role as the mark of understanding.

Next, I covered the candidate counterexamples that seem to involve abilities, but where the understanding attribution seems unwarranted. The first type were examples where the abilities are due to (i) a lucky shot, (ii) environmental or evidential luck, (iii) gettier luck, (iv) rote memorisation, (v) false beliefs, theories or idealisations, (vi) a short term, (vii) employing algorithms or models, or merely (viii) emulated. I argued that for each of these candidate counterexamples, the failure to counter my account comes from trying to warrant understanding through the lack of counterfactual acts, thus failing to show that the appropriate abilities are indeed present (or indeed lacking).

Lastly, I covered those examples where the abilities are due to (i) mimicking, (ii) reverse finks, (iii) external resources, (iv) a giant look-up table, or (v) blind rule-following. I argued that for each of these candidate counterexamples, the failure to counter my account comes from attributing the understanding or abilities to the wrong subject. To end, I discussed abilities that are (vii) derived from others (and why that shouldn't have us change the subject), as well as those that are (viii) lacking in

coherence (and why that should be addressed through characterising the object of understanding, rather than changing its mark, or its subject).

In discussing these candidate counterexamples and addressing them from within my presented account, I have further validated that account, showcased its strength (compared to others) and explained how it deals with many of the staple examples to be found in a variety of literatures.

PART II

CHARACTERISING EPISTEMIC SUBJECTS

PRELUDE 4

The Eye's Mind

An Aye-aye Doctor arrives at a tree in the middle of a forest. Mx. Spider calls out happily.

SPIDER: Thank you for coming, doctor! Did you find my web okay?
AYE-AYE: Yes, Mx. Spider, no issues there. I had written down the address.
SPIDER: Very good.
AYE-AYE: Now, what did you call me for?
SPIDER: I've been told that I should get my vision tested regularly. And I heard you were the best eye-doctor in the forest.
AYE-AYE: I don't know if I'm the best, but I'm certainly a certified eye/I-doctor.
SPIDER: Excellent! However, do I need to use my *eyes* for you to test my "eye-sight"? I tend to use my legs for eyes, you see.
AYE-AYE: No, "eye" is just a useful metaphor, really. I don't mind if you use your legs.
SPIDER: So, how does this work?
AYE-AYE: Well, I'll be presenting you with increasingly smaller insects, and I'd like you to tell me when and where you spot them.
SPIDER: Sounds easy enough. I hope I will impress you. You may not know this, but I'm known for my excellent vision. Give me the smallest one you've got.
AYE-AYE: Very well.

The Doctor uses his long finger to rummage around in the tree, and he takes out the smallest insect he can find. He then releases that insect into the air, near the spider. It doesn't take long until the insect flies into the spider's web.

SPIDER: There it is! I caught it in the corner of my eye, you see. And it seemed to go with relative ease.
AYE-AYE: Yes, it did. But I'm afraid I don't allow cheating.
SPIDER: Cheating? I didn't cheat.
AYE-AYE: You were clearly using that web of yours.

The Doctor climbs up the tree and takes away the spider from her web.

SPIDER: Oh! What? Did you just take me from my web without asking me?
AYE-AYE: Yes, I need to for the test. Is there anything wrong with that?
SPIDER: I suppose not, but I feel so naked all of a sudden. It would have been nice if you asked first. Why do you need to take it away anyway?

AYE-AYE: Well, I'm an eye/I-doctor, meaning I test both eyes and I's. Therefore I'm interested in how your I detects insects. It's important that we gauge *your* I's "eye-skills", not those of your web, you see. Start again.

The Doctor repeats the procedure and releases insects into the air near the spider.

SPIDER: It's like I have a phantom web. I still try to automatically reach for it even though I don't have it.

AYE-AYE: Don't worry about it. Just use your I to see!

SPIDER: But how? My web is missing.

AYE-AYE: Your web is not your I. Only you are you, that's tautologically true.

SPIDER: Sounds like you've just decided what's me and what isn't. That doesn't make it tautologically true, but circularly true.

AYE-AYE: Look, the web is clearly *external* to you. You can go wherever you want without it. You're doing it right now.

SPIDER: What is that supposed to prove?

AYE-AYE: It proves you aren't your web, you merely *use* your web. Now, just relax and tell me when and where you see the insect.

The Doctor releases another insect.

SPIDER: I can't do it without my web, I'm blind without it.

AYE-AYE: See, you agree with me. You say "*I'm* blind without *it*", which clearly shows you also separate your "I" from your web.

SPIDER: If my legs can be metaphorical eyes, why can't my web?

AYE-AYE: I don't mind our *eye* being metaphorical, but your *I* is not. I'm afraid *you* may simply be blind.

SPIDER: But that's unfair! With my web, I *can* see.

AYE-AYE: I can tell that you clearly don't know much about I's. "Seeing with your web," how funny! *You* can't detect anything with your web. Your web catches insects. And then what *you* do is detect web-vibrations, not insects.

SPIDER: What about the I that's me *and* my web? Together, we detect insects.

AYE-AYE: Don't be ridiculous, there's only one I in spider. And neither your eyes, nor your I, are detecting any insects.

SPIDER: Well, I'm afraid that's how I do things at home.

AYE-AYE: Then I'm afraid you are certified blind.

SPIDER: Well, if I'm blind, then Professor Raven can't fly, because she's a stemborg and one of her wings is entirely made out of twigs.

AYE-AYE: Totally different. I don't see how you can compare the two.

SPIDER: Then I think you should have your I's checked, mate.

Chapter 4

THE MARK OF EPISTEMIC SUBJECTHOOD & THE BOUNDARY PROBLEM

Understanding is always predicated on a subject. The one thing regarding understanding that was met with agreement by philosophers from the start was that it is pragmatic, meaning attributions of understanding stand or fall depending on the targeted subject. Therefore, that subject is, explicitly or implicitly, a key component in our considerations regarding "understanding." And yet the subject with understanding has not been considered in the literature with equal care as the mark of understanding has. The pragmatic nature (i.e. subject-dependence) of understanding has been discussed (and defended from assaults of absolute relativism) in the literature (and in Chapter 1 of this dissertation), but when it comes to discussing what makes for a relevant target, the literature on epistemology has little to offer. So far, it was often assumed that the targets of understanding attributions are (or should) always be human individuals, but there are cases that challenge that assumption. Many of our everyday and scientific abilities are more and more implemented by more than just individual humans (e.g. groups, artificial or coupled systems), and we need a way to conceptualise this with consistency and without an anthropocentric bias. Is there a systematic way to reveal what is required for subjecthood before we can attribute it with epistemic properties (such as understanding)? In other words, what is *the mark of epistemic subjecthood?* What is it that guides us in seeing an entity as a potential candidate for understanding attributions? Answering this question involves specifying what we find so philosophically or epistemically relevant about the epistemic subject. It involves clarifying how we are supposed to target, demarcate and conceptualise an epistemic subject.

To start, I would like to focus on our paradigmatic example, human individuals, and consider the how's and why's of its boundaries, and whether the skin or skull demarcation line is indeed fair or useful. When are things in the world appropriately part of the subject, and when are they merely the environmental embedding of that subject? In other words, where do we draw the line between what extends an individual and what embeds one? While the answer may change depending on what one is interested in, I will argue that a good guideline is to let the boundaries be dictated by what implements a coherent and persisting epistemic agent. This means that, as a mark of epistemic subjecthood, I will defend the interpretationist approach, and more particularly the epistemic stance (the intentional stance with an epistemic focus). The epistemic stance is the instrumental strategy of interpreting behaviour by treating it as if the entity were governed by beliefs, epistemic aims and epistemic tactics (as well as any other intentions that play a supporting role). Having defended the epistemic stance as the mark of epistemic subjecthood, I will argue that if an entity, composed of more

than just a human individual, can grant us explanatory or predictive powers through the epistemic stance, then taking advantage of this power is not only warranted and fruitful, but consistent with our best conceptualisations of individuals. In that case, we are dealing with an extended epistemic agent. To end, I will discuss 7 different cases to showcase what gets extended in extended understanders and how.

4.1 The Value of a Mark

In this chapter, I shall argue, with inspiration drawn from the interpretationists, that epistemic subjecthood boils down to a virtual agent being an explanatory and/or predictive concept in the context of epistemology. But before I propose a unifying mark of subjecthood, it will be worthwhile to consider what would make a mark valuable in the first place.

The Value of a Mark of Epistemic Subjecthood

Up until now, the emphasis of both the literature and this dissertation has been on the mark of understanding. But whenever we make understanding attributions, there is always a target for those attributions. If we wish to take those attributions seriously, we don't just need to be able to mark the attribute, but also specify its target, the epistemic subject. So far, it's often been assumed that the target of understanding attributions are (or should) always be human individuals, and that it will furthermore be clear (in each case) which human individual is the appropriate target. In this dissertation I have been defending an ability approach to that mark. Human individuals can certainly display some impressive abilities and it is very commonplace to, based on these abilities, attribute understanding to the targeted human individual. For instance, Inga understands why a theorem is true because she can consistently work out its mathematical proofs, can explain the outline of the proof to a non-expert, show a reductio ad absurdum if the theorem were untrue, etc. Furthermore she can accomplish all these things without any "outside" help. She works out all the problems by herself and then relays them to us. Otto, on the other hand, due to problems with his working memory, can't do any of that. Based on the ability-approach to understanding, it seems fair to say Inga's understanding is superior to that of Otto.

As long as it is indeed easy to target the appropriate individual human, it doesn't seem like we would have to rely on a philosophical account of epistemic subjecthood just to be able to target the appropriate entity. For many of our everyday understanding attributions that would appear to be all there is to it. But there are cases where our intuitions seem to differ or where we, in the absence of a concept of epistemic subjecthood, have a hard time motivating who does or does not warrant an

understanding attribution and why. For instance, was it fair to target Otto by himself? What if all he needs is a notebook to do the same things Inga does without one? Many of the abilities in the sciences and everyday life are more and more frequently including things from our environment (e.g. pen and paper, smartphones, computers, calculators, physical models) to achieve them, and we need a way to conceptualise this.[127] Furthermore, everyday language is full of understanding attributions where the targets are entities other than human individuals, and it is not always clear whether this is meant as a literal or metaphorical attribution (and if the latter, to which degree). We can say things like "Nestlé understands that it can make more money if they pretend they don't know about their child labour", "the Flemish government doesn't understand the severity of the climate crisis" or "CERN is gaining an understanding of the physical laws governing the behaviour of matter." In each of these, the target is a group. We can also say things like "Google Maps doesn't seem understand that I'm okay with walking over 15 minutes", "the chess program understood what I was trying to do, and foiled my schemes" or "Even Coq [an interactive theorem prover] understands that widened iteration terminates" (Leroy, 2014, slide 24). In each of these, the target of the understanding attribution was an artificial software program. Are these understanding attributions to groups or artificial systems merely *metaphorical?* Are they always? And are they *different* from understanding attributions made to human individuals? If so, to what extent are they different and why is that difference relevant for understanding attributions? In the absence of a concept of epistemic subjecthood, we are relying on intuition only, and intuitions can be conflicting, misleading, naive or inconsistent.

For attributions to be explanatorily meaningful, there needs to be a (relatively) persisting and coherent target, an epistemic subject, to which they apply. One of the reasons why we make attributions is because it helps us interact with entities, and explains something about what sort of entity we are interacting with. "Inga understands basic calculus" is a helpful assessment because it (a) identifies and targets an entity which persists over time, namely Inga, and (b) it reveals something about that persisting entity, namely that she is able to answer questions or solve problems in basic calculus at the opportune time (e.g. when comfortable and prompted). But if such abilities can be displayed by non-human entities, may not the practice of attributing understanding to such entities be equally helpful? Only with a concept of epistemic subjecthood (along with the knowledge of its limits and pitfalls) can we be explanatory, while also being consistent and conscientious about the concepts we use. Having to mark epistemic subjecthood thereby challenges the idea that only human individuals

[127] It is of course possible to simply see certain resources as not relevant enough to be worth mentioning. If some resources are very commonplace (for a particular context), they can be considered as mere background conditions. My account can incorporate this move through the economy weights (see Section 2.2). Nevertheless, this does not help us demarcate the epistemic subject, so we are still in need of a mark.

are appropriate targets and forces us to clarify what belongs to an entity, and what doesn't, as well how we should conceptualise how things in the world integrate into one subject. In short, we need a mark of epistemic subjecthood.

And lastly, having a mark of epistemic subjecthood allows us to consider understanding with a richer picture in general. As was clear from the objections in Chapter 3, Section 3, the absence of a mark of epistemic subjecthood may contaminate how we deal with the mark of understanding. For instance, the use of external resources was deemed a cause to discredit abilities, the use of a giant look-up table was deemed an invite to look behind the abilities, blind rule following seemed to lead to importance of excluding exact calculation. And yet, in all cases of Section 3.3, understanding got denied simply because the wrong entity was targeted, not because the wrong sort of mark was adhered to. This is a pitfall that can easily be avoided by combining the mark of understanding with a mark of epistemic subjecthood.

The Mark Question

If having a mark of epistemic subjecthood is valuable, then we need to have a discussion about the question of what it is that marks an epistemic subject and why. Answering such a question involves a conceptualisation of what makes up a subject and what doesn't, drawing a line between what lies within the boundaries of the subject and what doesn't, as well as why we draw that line where we do. So how do we go about answering such a question? The value, use and intended aims of the concept of epistemic subjecthood can help us indicate some of its requirements.

Firstly, if an epistemic subject is supposed to be the target for attributions of epistemic predicates (understanding, knowledge, etc), then the concept of an epistemic subject needs to make sure that such attributions are explanatorily meaningful. The first point here is that these attributions require an entity that is relatively constant or persisting enough to benefit from predicate-attribution. This entails that the entity is relatively *cohesive* (physically or functionally connected in producing the epistemic acts). Without (relative) cohesion, it is unclear what makes it one subject, instead of many. For how would it make sense to make attributions if nothing keeps the targeted parts of the world together in any way.[128] Next, it means that the entity is relatively *coherent* (singular in its epistemic

[128] This does not have to mean they have to be literally glued together. If a brain was connected to a body via Wifi to produce its acts, then the brain's location and freedom of movement independent of the body would not make it distinct, since they functionally act as one. Yet if a brain was inside a body, but severed of all its connections with it, then the brain and body would not both be functionally involved in the production of epistemic acts, and therefore hard to see as one. More on that later.

identity) over time. Without (relative) coherence, it is unclear what makes the attributions befitting for the same subject. For how would it make sense to make attributions to an entity if nothing keeps the several attributions together in any way. Even if an entity is physically and functionally connected, if its identity is disjointed, then its subjecthood will be as well.[129] Furthermore, this cohesiveness and/or coherence must be relatively persisting. Without persistence, the appropriateness of the attribution would be as changeable as the entity it is supposed to be attributed to. And if (epistemic) subjecthood falls apart, so does the target of our evaluations. In other words, if there is no (relatively) persisting entity with cohesiveness and coherence, then there is no point to the attribution.[130] We need something that ties things together in the world such that they make up one subject, in time, in identity and in acts. In other words, we need a principle of composition or integration.

Secondly, if there are target entities that are attributed with epistemic predicates, but some of them are metaphorical and others are "real," then characterising the subject will involve being able to distinguish between those which are mere metaphors and those which have a stronger metaphysical weight to them. Epistemic subjecthood is a way in which metaphorical and real attributions can be distinguished.[131]

Lastly, having in the earlier chapters characterised understanding in a specific way (i.e. marking understanding through the contextually appropriate abilities), it would behove us to characterise the epistemic subject in a way that coheres with that characterisation. Being able to do so would furthermore bolster the strength of both characterisations. This means that epistemic subjecthood should at least be able to conceive of an ability-based epistemic predicate, and not fall into similar traps that we criticized earlier. If we were to mark epistemic subjects by the presence of something in an indiscernible mental sphere, then we would have equal difficulties in discerning that mark in others or knowing which aspect is valuable to target in the first place. So any principle of composition that

[129] For example: Dr Jekyll and Mr Hyde are two subjects in the same body. If they were each given an exam, their scores (and the explanatory power of those scores) might not align, even if their body does - therefore, it makes more sense to give attributions to each separately rather than to both. The predicate of understanding may explain Dr Jekyll but not Mr Hyde or vice versa. Only considered separately will they have a coherent epistemic profile. Whatever this profile consists of is something the concept of epistemic subjecthood must supply.

[130] The simplest example is that the death of a subject disbands both its persisting subjecthood and the epistemic attributions predicated on it.

[131] Of course, epistemic subjecthood may not suffice to distinguish them fully. Chapters 5 and 6 are devoted to further problems in attributing understanding to epistemic subjects (beyond the mark of epistemic subjecthood), namely that of reducibility and regress respectively. To ensure that the target entity is the most appropriate target of the understanding attribution, these problems will also need to be addressed. Nonetheless, as will become clear, these further issues do not challenge the mark of epistemic subjecthood, but merely extend it.

ties things together in the world should put its premium on acts, and not on something *behind* them - unless it does so instrumentally. Fortunately, this is exactly what interpretationism does.

An already ubiquitous characterisation of subjecthood is of the subject as an *agent*. An agent is cohesive, coherent and persisting entity attributed with certain properties or states (e.g. beliefs, intentions and rationality), some of which are epistemic. The interpretationist approach to agency, in particular, is promising because it is a systems theory based on the explanatory and predictive power of a virtual entity (i.e. the agent) and the properties it postulates (i.e. its beliefs, intentions and rationality) in order to account for the entity's behaviour. I will be relying on this approach to conceptualise the epistemic subject in a way that helps us deal with subject demarcation - not only in human individuals, but also in coupled systems (see Section 4.4) and non-human entities such as groups (see Chapter 5) or artificial systems (see Chapter 6).

The Environment Question

The flipside of the mark question (i.e. conceptualising what belongs to the subject) is the environment question (i.e. conceptualising what doesn't). This question may be of equal importance, because subjecthood and understanding are always at least partly due to the environment. It is just that we are not always equally interested in the contribution of that environment, especially if it can be kept constant. For instance: when we target Inga and Otto, as human individuals (meaning we demarcate them at their skull), it is useful to know what they can and cannot do without any further resources. Nonetheless, the value of isolating the human individual is not ubiquitous and there are dangers to keeping the environment metaphysically constant. Dangers that are similar to assuming everyone in society has equal opportunities, and that if a certain group of people don't score as high in life, it must because there's something wrong with them (as opposed to their environment). Not acknowledging differences in environments can lead to massively distorted views about the differences in subjects. This is true even under an ability approach.

> "Much of cognitive science is an attribution problem. We wish to make assertions about the nature of cognitive processes that we cannot, in general, observe directly. So we make inferences on the basis of indirect evidence instead, and attribute to intelligent systems a set of structures and processes that could have produced the observed evidence. This is a venerable research strategy, and I have no objection to it in principle. However, failing to recognize the cultural nature of cognitive processes can lead to a misidentification of the

> boundaries of the system that produced the evidence of intelligence." (Hutchins, 1995, p. 355-356)

Scientists (be they physicists, meteorologists or mathematicians) do not work in a vacuum, nor are they expected to. So if we wish to be able to account for their environmentally supported abilities, we need a way to address that, conceptually. I'll refer to this denial of the contribution of the environment as the *shrinking problem*.[132]

Nonetheless, there are also dangers to assimilating too much of the environment. Just because a weather forecaster looked at the clouds to predict the upcoming rain does not make the clouds a part of her as an entity capable of predicting the weather. Similarly, just because a mathematician uses a chair while proving, that does not make the chair part of her. I'll refer to this excessive focus on the contribution of the environment (however trivial) as *the bloating problem*. So not everything that played a role in accomplishing abilities need be included as part of the ability-having subject. We have to draw the line somewhere between input and system. Nonetheless, if we wish to be able to account for environmentally supported abilities, we need a way to address that, conceptually.

To take seriously the role of what is external to human individuals, there are two possible routes: saying these abilities are embedded or extended. In both approaches, we need to take seriously that abilities (or cognition, or the mind) are not isolated inside the head of individuals. In the embedded approach, we do this by taking seriously the role of the environment required for the individual to display her abilities, or cognition or mind (a course taken by, for example, Putnam and Burge). In the extended approach, one needs to take seriously that the role of the environment may not always be significantly different to the role of the human individual to warrant the traditional dividing border between the two (a course taken by, for example, Clarke and Hutchins). In the extended approach, we need to have a way to stop the bloating problem. In the embedding approach, we need to have a way to stop the shrinking problem. But when do we embed the abilities of the individual and when do we extend the subject? It is important that the mark of epistemic subjecthood addresses why it is useful to draw the border where it does, on pain of needless bloating or shrinking.

[132] The shrinking problem can be found even below the level of skin. Consider this quote by Andy Clark (cited in Tollefsen, 2006): "Go into the head in search of the real self and you risk cutting the cognitive cake ever thinner, until the self vanishes from your grasp. For there is no single circuit in there that makes the decisions, that does the knowing, or that is in any clear sense the seat of the self. At any given moment, lots of neural circuits (but not all) are in play. The mix varies across time and task, as does the mix between bodily and neural activity and all those profoundly participant non-biological props and aids." (p. 146)

What we need is a mark of epistemic subjecthood that would help us target a relatively persisting target for who the attributions of epistemic properties (e.g. understanding, beliefs, etc) would be explanatory or predictive. This entails that the entity needs to be relatively cohesive (physically or functionally connected in producing the epistemic acts) and relatively coherent (singular in its epistemic identity). Furthermore, it needs to withhold needless shrinking or bloating beyond what would be explanatory.

4.2 The Mark of an Epistemic Subject

So the concept of epistemic subjecthood is a way of targeting a (relatively) persisting, cohesive and coherent entity for which it makes sense to engage in attributions of understanding. But what are good candidates for the concept of epistemic subjecthood? Which approach allows us to tie together things in the world into one subject (ready for us to consider its abilities and understanding) and conceptualise its internal workings (on the basis of its acts)? Before I offer up my arguments in favour of the epistemic stance (the intentional stance with an epistemic focus - to be explained in Section 4.3), I would like to consider two different approaches to demarcating a subject. I have never seen either of these get explicitly defended as a proper demarcation-criteria (they are too naive for that), but they both carry strongly intuitive tenets which I have seen in arguments for or against subject-demarcation. Therefore, considering them will help clarify the strengths, weaknesses and pitfalls of certain intuitive arguments we will encounter, as well as how the epistemic stance deals with them.

Demarcating at Will

The first approach is possibly the easiest approach to demarcating a subject: just doing so. I will refer to this as the border-targeting approach.

> *Border-targeting approach*: Something belongs to the subject if it is targeted as part of the subject.

In the border-targeting approach, one simply decides which physical boundaries are relevant, as a starting point, to then determine which acts or abilities the target entity supports, and subsequently gauge whether they are sufficiently appropriate for understanding attributions. While it is a beautifully simple approach, there are a couple of issues with it.

Firstly, the border-targeting approach does not posit anything explanatory, such as a conceptualisation of the internal workings of the entity, why it acted the way it did or whether it will

act the same way in the future. It demarcates, but does not explain. Secondly, there's no systematic targeting of things in the world, and if there is, it is not made explicit. This approach does not help clarify which parts in the world belong together (or why) because it merely asks you to choose which things you would like to consider *as* together. Therefore, the border-targeting approach does not offer a demarcation-principle so much as a demarcation-license.

Thirdly, under this approach we may lack a persisting and coherent entity. If we target things blindly, there is no guarantee that the behavioural profile of that combined entity will cohere. In fact, its output may consist of acts that do not inform one another and may even go as far as to contradict each other. When I target my little sister, my calculator and a research-centre, so I can then try to find out the position of the combined entity on ghosts, I may find that the relevant acts I find do not cohere and possibly even contradict each other (e.g. I may find assertions that ghosts are real, that the local parish is haunted, an error message, and that any belief in the supernatural is unfounded) or its views simply do not inform one another (e.g. I may find actions of avoiding the local parish and campaigns to dissuade people from their fears of it). Some of these acts cohere with each other better than others, and this may be an indication that drawing a different border, based on some other criteria, would be more appropriate. For instance, the above example would be better explained if we paired up certain acts with certain entities (e.g. my calculator revealed an error message with no understanding or position on ghosts at all, my little sister asserts that ghosts are real and avoids the parish, while the local research-centre asserts that any belief in the supernatural is unfounded and attempts to persuade people to give up their fears), but then we are moving away from our border-targeting approach and towards another (as of yet implicit) criterium. To defend border-targeting, you could insist that the combined entity is simply a single entity that is not perfectly rational or coherent. This is a possibility that should not be excluded (even individuals aren't perfectly rational or coherent all the time), but this explanatory hurdle would need to be compensated in some way by the explanatory power of continuing to treat the combined entity as one, for which I see no evidence.

This is not to say that this approach is never useful. Sometimes we do start from physical boundaries, simply because those boundaries themselves are more important to our purposes than the fact that there would be a coherent persisting entity we could call a subject. It can be especially relevant to target things that are often found together to find out what their behavioural profile will be in those multiple situations where they are together. This is the case, for instance, with certain modular entities (e.g. "How does the Fairphone perform with its new battery?"), specific groups (e.g. "How well behaved is the new 6th grade class?"), etc. But most cases of border-targeting will be nonsensical

because the combined entity produce coherent output that persists over time (e.g. it is nonsensical to ask "What is the behavioural profile of my little sister's frontal lobe + this tea mug + that computer?"). Furthermore, the border-targeting approach fails to account for the fact that an entity can persist coherently over time even as it changes its physical components. Consider entities such as companies (e.g. "Has Apple learned its lesson since last year or is still using sweatshops?") or sport-teams (e.g. "How is your favourite team doing this year?"). Even though each of them can be considered as a persisting entity in some way (which is why they carry the same label over time), what persists in each of them is not the physical things that constitute them at any particular moment. The Fairphone's parts can be replaced with new ones, a company's employees and board of directors can be renewed and a sports-team's players can be replaced. So if the physical components are not necessarily what persists, how can we define subjects through targeting those components directly?

It may still seem tempting to defend the border-targeting approach because human individuals are targeted via presumed borders. Human individuals, whatever their internal make-up and relation to their environment, tend to take whatever they have underneath their skin/skull with them everywhere they go. So it seems particularly relevant to know what the behavioural profile is of that portable and persisting entity, bounded by skin/skull. I will grant that this is an intuitive place to start and that it proposes a clear demarcation criterium. It tells us exactly where the subject begins and ends, it does so in a way that can be systematically applied to different situations or entities and it does so in a way that fits with what we take to be a paradigmatic example of an epistemic subject: a human individual. Nonetheless, it doesn't contribute any useful conceptualisation of what goes on inside a subject, nor does it tell us why the skin or skull should be the relevant barrier. That problem becomes more pressing when there are arguments that challenge the idea that the skin or skull are a relevant boundary. For instance: consider a subject possessing a mathematical proof. A picture of a proof may be hidden behind the occipital lobe, but its passive presence there is not *sufficient* to warrant its inclusion as part of the subject, especially not as an internal representation (seen instrumentally). After all, its passive presence will play no role in our conceptual characterisation of that subject. But the problem is not just passivity alone. There are multiple examples in the epistemology literature of a brain tumour or lesion being considered as separate from the (epistemic) subject. Consider this comment:

> "Brain lesions (...) (of the proposed sort) do not count as part of agent character, and precisely because they are not well enough integrated with other of S's cognitive dispositions." (Breyer & Greco, 2008, p. 175)

So even if something can be seen as a functional part of a subject (affecting its decisions and abilities), there can be cause for concern whether these acts are sufficiently integrated into (or cohering with) the rest of the subject to warrant treating it as one entity. So the skin and skull boundary doesn't guarantee functional cohesion or coherence.

Neither would the presence inside the skull be *necessary* to consider it as part of the subject. If part of the brain were taken outside of a mathematician, its neurons kept alive and its signals connected (via high-speed WiFi) to the original part inside the skull, then does it really matter that part of the brain is not physically "inside"?[133] Now consider the reverse: If a mathematician can achieve the relevant abilities, but only with the use of artifacts, then we need to be able to recognise that these abilities are still relevant even if the realising base for them stretches outside of the skull of the individual. The skull and skin characterisation may be a useful one because human beings are persisting entities that take their skull and skin (along with everything in it) everywhere they go, so targeting those things as a single entity of which we may determine its behavioural profile is a useful thing to do. But is it really the skin or skull that is doing the heavy lifting in demarcation, or is it just a frequently useful boundary? The previous examples suggest that it is merely the latter.

Demarcating from Acts

In the second approach that I will briefly consider, one starts from a particular set of acts (whether they form an ability or abilities, not) to subsequently determine the physical boundaries of what realises those acts. I will refer to this as the act-realising approach.

Act-realising approach: Something belongs to the subject if it realises the targeted act(s).

The first task then becomes deciding which act or acts (or which ability or abilities, made up of acts) one is interested in. Having found an appropriate set of acts, we can look for all of the physical bits and bobs in the world that helped realise it, by virtue of playing a functional role in bringing the act about. This approach also faces a couple of issues.

Firstly, the act-realising approach does not posit anything explanatory, such as a conceptualisation of internal workings of the entity, why it acted the way that it did, or whether it will act the same way in the future or not. It demarcates, but does not explain. Furthermore, it leads to the possibility of needless bloating. The bloating problem is partly due to the fact that there is no conceptualisation of

[133] For a number of thought-experiments related to these, see (Dennett, 1978b).

the internal workings of a subject, so everything that was used will be counted as part of the subject. It doesn't specify the difference between system and input or system and incidental environment. It is important that the mark of epistemic subjecthood addresses why it is useful to draw the border where it is drawn, on pain of needless bloating (or shrinking).

Secondly, if we start with choosing a particular set of acts or abilities, the criteria for which acts or which abilities belong together is entirely left open. They may be utterly random, or, worse, inconsistent. For instance, they may include both the endorsements of p and the endorsement of not p. Now, it may be fair to assume that someone using the act-realising approach simply won't start from an inconsistent set of acts, but the point is that the reason and method for selecting acts remains entirely implicit or hidden. As such, the realiser approach does not offer any explanations or criteria of relevance for which particular sets belong together, or why they would be a relevant set to demarcate the subject with understanding from. This entails that they may constitute a collection we would not consider as acts of a (single) subject (e.g. asserting that ghosts exist, and campaigning that there's no such thing as ghosts). The approach is, in short, question-begging about the relevance of the acts one starts with.

Thirdly, under this approach, we may lack a persisting and cohesive entity. There is no guarantee that the realising base of the appropriate acts has any cohesive bonds, let alone persisting ones. We could form a coupled system constituted by anything we want, just to accumulate its abilities. This becomes especially pressing if we are targeting more than a single type of act. If the realising base for the relevant type of acts you've been targeting (e.g. predicting the weather for tomorrow, calculating the distance between A and B, and formulating an evacuation plan) turn out to be realised by Ororo, your computer and the local crisis-manager respectively, with no physical bonds or even interaction between any of these things (let alone any persistence in those bonds), then we have to question why it is worthwhile tying them together into one subject. This becomes especially clear if the abilities of one sub-entity would be relevant to another. The local crisis-manager may be able to form evacuation plans for any number of emergencies, and Ororo may predict which one will happen tomorrow. But without any interaction between the two, one will not inform the other. Their abilities as a coupled system are nothing beyond a mere trivial and disjointed summation of their individual abilities. In other words, there would be a lack of cohesion. Furthermore, the problem would remain if they did interact, but if their interaction did not result in any coherence. If Ororo predicted a storm and told the local crisis-manager, who prepares for a flood instead, then the coupling of their acts remains a trivial and disjointed summation. We might be better off linking each type of act to a separately

demarcated entity (e.g. Ororo predicts a storm, your computer calculates the distance between A and B, and the local crisis-manager is able to put together an evacuation plan for a flood). The acts may together mark some understanding, but there's no particular entity that marks an epistemic subject to whom the understanding can be attributed.

In short, if we don't have any further criteria for what makes a set of acts worthwhile to select together under the act-realising approach (and I don't know of any such candidate theories), we may simply have a fragmented conglomerate of acts realised by disjointed things in the world. Of course, this is not to say that it can never be of interest to know which things in the world realise which random set of acts (regardless of whether the acts cohere or the realising bases "cohese"). Examples include questions like "how did this accident occur?" or "why are there storms?", etc. For certain purposes, it is indeed more relevant to know what the realisation of a conglomerate of acts (or events) is, but the concept of subjecthood was supposed to help us reveal a cohesive and persisting entity of some kind that can be stably attributed with a coherent quality of understanding (a scope of sensitive, stable and efficient performances), and the act-realising approach does not offer any tools or criteria to do so.

So neither the act-realising approach or the border-targeting approach constrain the set of their targets in any usefully systematic way. Both start from a set of targets in the world (either a set of things or a set of acts), with the criteria for why it belongs to the set left open. This move is fair, of course, in limited cases, but it doesn't shed light on why we consider certain entities as subjects, epistemic or otherwise. For epistemic subjects, something needs to tie together the things and acts so one coherent epistemic profile is found. (Unless, as mentioned, you are not interested in a coherent subject, but merely in what realises a particular set of acts.) Furthermore, something needs to tie together the physical things, so one cohesive entity is targeted. (Unless, once again, you are not interested in a cohesive subject, but merely in what is realised by a particular set of physical things.) For this, we need something more systematic, both to determine the sets of acts we are to consider and to determine which physical boundaries we are to draw.

Cognitive Character

In the epistemology literature, there is a helpful concept we can find to mark epistemic subjecthood, namely that of a *cognitive character*. This concept helps clarify what (be it a belief, an act, an ability) belongs to the epistemic subject. At least part of the reason why Pritchard (2010) uses the term "cognitive ability" is to make sure that the ability in question actually belongs to the epistemic subject in question, or to the "cognitive character" as he (and others) call it. According to Pritchard (2010) "an

agent's cognitive character is her integrated web of stable and reliable belief-forming processes." (p. 136) This rough description already seems to indicate that trivial summation of acts is indeed insufficient, because without coherence or a functional link between the acts (or between the beliefs, intentions, etc we can derive from acts), there is no evidence of *integration*. The idea of integration into a cognitive character sounds like a plausible candidate, but if we are to adopt it as the mark of epistemic subjecthood, we need to know more about what makes up a cognitive character and how do we determine what is part of it. For that, there are several suggestions.

The first is based on the concept of *reflective endorsement*, meaning something (e.g. an intention) belongs to the subject if it is reflectively endorsed. Here, the beliefs (or acts) of the subject are those that the subject identifies with. (Breyer & Greco, 2008) But this leads to a couple of problems. Firstly, because the emphasis is on *reflective* endorsement, it moves the mark of epistemic subjecthood *behind* the acts, into a realm where no one, but the subject itself, can peer. We are already familiar with the problems of this move from Chapter 1. Additionally, the requirement of reflective endorsement also entails that Jake may identify as a non-smoker, even while constantly smoking. Conversely, it also requires that Jake is not an alcoholic because he does not identify as such. The subject must be both aware of all of its (subject-)properties, and endorse them. This while the properties we regularly attribute to subjects may go unnoticed, or may be denied. (While it is not trivial that Jake may endorse that he wants to stop smoking, if his everyday behaviour conflicts with that endorsement, then equating Jake only with the acts and beliefs he identifies with is merely ignoring that conflict.)

The second suggestion, supposedly analogues to Fischer and Ravizza's (1998) account of moral responsibility, focuses on the notion of *taking responsibility*. Something belongs to the subject if the subject takes responsibility for it. (Breyer & Greco, 2008) While taking responsibility can be a pathway to integration, taking responsibility alone does not guarantee it. I may choose to take responsibility for the colonial acts of Belgium (and not only my current unwarranted privileges that come with it), but that does not make them *my* acts.

The third suggestion, based on Ekstrom's (2005), is a structuralist one that Breyer and Greco (2008) defend from a reliabilist perspective.[134] They believe Ekstrom can be read in two possible ways. The first is as a coherentist account, where "cognitive integration is a function of coherence among beliefs,

[134] Reliabilists, in epistemology, mark knowledge by putting the premium on processes that reliably produce the truth. (Goldman & Beddor, 2016) They approach things from an external point of view and are therefore more complementary to ability-oriented accounts of understanding than internalist ones.

and belief ownership is understood in terms of membership in this integrated structure" (p. 182 - 183). This interpretation would satisfy our requirement that acts should cohere with one another through that "function of coherence". The second is as a dispositional account, where "cognitive integration (...) [results] from the cooperative causal interaction of relevant cognitive dispositions." (p. 183). Here, a belief belongs to a subject S, "insofar as it is a product of S's cognitive character, where cognitive character is a causally integrated system of cognitive dispositions that are themselves aimed at truth." (p. 183) This would satisfy the requirement for cohesive bonds through causal integration. I believe both the focus on coherence of beliefs as well as the causal cohesion behind dispositions are appropriate, but we still lack guidance in what ties these beliefs and/or dispositions together.

One last suggestion comes from Pritchard (2010). He argues that one can bypass the discussion on the nature of cognitive character "by simply focussing on the question of whether we would treat the agent's cognitive success as appropriately creditable to her cognitive agency." (p. 137) As an example, Pritchard asks us to consider Alvin. Alvin is someone with a brain lesion, and thanks to that brain lesion, he successfully and reliably forms true beliefs. Nonetheless, it is intuitive that in spite of the reliable true beliefs, Alvin does not have knowledge (or understanding, for that matter). The primary reason for this is that those beliefs clearly have nothing to do with Alvin, and everything to do with his brain lesion. The beliefs we may infer are not Alvin's beliefs, but those of his brain lesion. His example is not unlike that of Henrietta and System Hyde (see Section 3.3, example v), and I too appealed to the difference in agency between the two. I believe this move is a fruitful one, but it does not yet bring us to our destination. This appeal to agency does not bypass the discussion of epistemic subjecthood, it merely shifts it to a more intuitive concept, namely that of cognitive (or epistemic) agency, and I will now show that it is worthwhile to further clarify the nature of that agency.

Interpretationist Demarcation

Some features which are routinely associated with the concept of subjecthood are beliefs, ideas, intentions, etc. These features are classic components of agency. And even if agency is not directly about demarcation, we will see that it does conceptualise the subject in an explanatory way that also helps with demarcation. The approach to agency I will focus on here is one we have already discussed in Section 1.4, namely Daniel Dennett's (e.g. 1990, 2009) interpretationist approach through the intentional stance. Here's a reminder of what it entails:

> "Anything that is usefully and voluminously predictable from the intentional stance is, by definition, an intentional system. The intentional stance is the strategy of interpreting the

> behavior of an entity (person, animal, artifact, whatever) by treating it as if it were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires.'" (Dennett, 2009, p. 339)

According to Dennett, the intentional stance is an innate capacity (as opposed to an academic theory) to interpret an entity as being governed by beliefs, intentions, and "and enough rationality to do what it ought to do given those beliefs and desires." (p. 340). The sole justification for considering an entity as an agent is the efficacy of the stance in predicting or explaining its behaviour that way (regardless of how it is realised physically), so there is no difference between a "real" and an "as if" agent, and no dividing line between the two. (Dennett, 2009) It is a normative stance in that the interpretation depends on what the agent ought to do (rationally), and requires a holistic approach in that the success of the stance lies not in pairing up the components of the stance with particular behaviours, but in how well the agent-package predicts or explains the entity's behaviours overall. (Dennett, 1990)

The intentional stance is a more promising candidate for marking an epistemic subject because it conceptualises the entity in a way that is useful for attributions. This is because it targets a coherent persisting target. After all, the stance would not be explanatory or predictive unless there was something coherent and cohesive it could explain or predict. Furthermore, the intentional stance also allows us to also (indirectly) demarcate. This is because it allows us to focus on the realising base of those acts that reveal the agent's coherent features. This means we have a demarcation criterion which tells us both what belongs to a subject virtually (i.e., the components of the intentional stance) and physically (i.e., the realising base for those components). Therefore, the interpretationist approach is a worthwhile candidate and roughly the strategy I will be following here (with some further caveats, to be explained in Chapters 5 and 6). But the intentional stance only tells us something about subjects, not *epistemic* subjects. For that, we will need the slightly more focused concept of *epistemic agency*.

4.3 Defending the Epistemic Stance

If the mark of epistemic subjecthood is epistemic agency, then our question shifts to what it is that warrants anyone (or anything) being an epistemic agent. My contention is that it is nothing more or less than being a successful target of what I call the epistemic stance. This approach to epistemic agency is predominantly lifted from Dennett's interpretationist conception of agents through the intentional stance, except with a focus on the epistemically relevant properties. To distinguish the

two, I shall refer to *the epistemic stance* instead, but the difference with the intentional stance is not one of kind, but one of focus.

Features of the Epistemic Stance

The epistemic stance is the strategy of interpreting behaviour by treating it as if it were governed by epistemic aims (i.e. the kind of results that an epistemic practice values), epistemic tactics (i.e. any serious systematic attempt to get closer to an epistemic result), and beliefs (which are always epistemic, so don't need the modifier), as well as any other intentions and tactics that play a supporting role in the epistemic agency. What counts as epistemic, and what doesn't, won't always be neatly distinguishable. Keeping a room at a particular temperature is not an epistemic aim. Applying for funding or drinking tea to help with concentration is only on the very fringes of epistemic aims or tactics, whereas breaking down the problem into parts or writing down a mnemonic device to help navigate a search-space is properly epistemic. I shall assume that the reader and I share similar enough intuitions about what counts as properly epistemic and what doesn't at all, and leave the grey areas for what they are.

Crucially for our purposes, this approach to the concept of epistemic agency, much like our approach to the concept of understanding, is act-based. Furthermore, when it comes to abilities and agency, one tends to invite the other. It would be unlikely that an entity displaying a set of abilities does not allow for an explanatory story of beliefs, epistemic aims, and tactics accounting for its actions. And, conversely, it would be unlikely (though not as unlikely) to find an entity with no appropriate abilities whatsoever, but where an interpretation of beliefs, epistemic aims, and tactics does, nonetheless, have explanatory value. The strength of the intentional stance doesn't lie in revealing what's "really behind" the acts, but in exploiting a kind of pattern in acts to explain and/or predict further ones.

The components of an epistemic agent are beliefs, epistemic intentions and epistemic tactics. In an earlier paper, I have referred to a subject having a background/skill corpus (Delarivière, Frans & Van Kerkhove, 2017), but the components of epistemic agenthood as conceived through the epistemic stance (beliefs, epistemic intentions, rational tactics) is a richer way of fleshing out the same idea. The epistemic stance is an instrumental abstraction. It is instrumental in that the sole justification for interpreting an entity as an epistemic agent is the explanatory and predictive success of that interpretation. If seeing an entity *as* governed by epistemic attributes (beliefs, epistemic aims and

tactics) has explanatory or predictive power, then, by definition, that entity is an epistemic agent.[135] Like the intentional stance, it is normative and the normative standards of an epistemic practice are brought to bear in the interpretation. What an epistemic agent ought to do or ought to believe, can only be interpreted given the standards of the epistemic practice to which the epistemic agent belongs.[136]

The epistemic stance reveals an instrumental postulate, the epistemic agent. And this epistemic agency is revealed by a *macro-systematicity.* I will explain what systematicity is and what makes it macro in the subsequent subsections. But in short, *systematicity* is a pattern (i.e. the epistemic agent) that a theory (the epistemic stance) can predict or explain. Like the intentional stance, the virtual pattern it reveals is not atomistic, in the sense that its components (e.g. belief p) correspond directly to individual acts (e.g. endorsing p), like it would under behaviourism. Instead it is holistic, meaning that components can only predict or explain the behaviour *as a whole.* Furthermore, its systematicity can only be detected at a higher, *macro,* level. Like the intentional stance, the epistemic stance makes no dictates on what it is a pattern of, so it does not rely on there being a direct correspondence between our ascriptions of beliefs, aims and tactics and some structure in the brain. The components of the epistemic stance are virtual, not physical. They track something salient in the entity, not because there must be a literal implementation of its structures in that entity, but because it is instrumental in explaining or predicting the entity's *systematicity*. The postulate (i.e. the epistemic agent) revealed by the epistemic stance is thus a salient virtual macro-systematicity. What exactly does that entail? I will address the macro-level and its systematicity in turn.

Macroscopes & Levels of Explanation

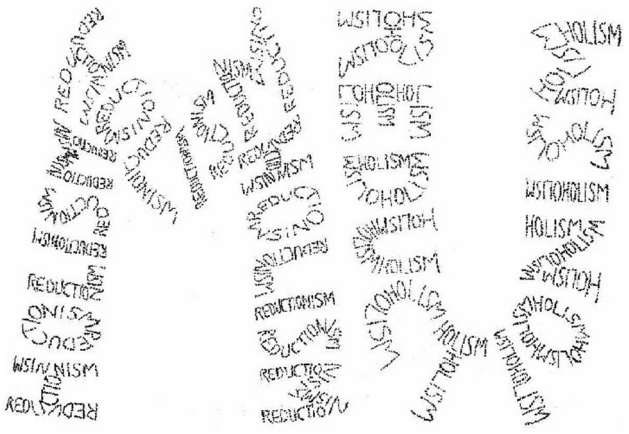
The epistemic stance aims to predict or explain. Nothing would allow us to predict or explain happenings in the world more precisely than discerning and utilising precise laws of nature (through what Dennett calls the physical stance). At a very small level, a micro-level, there may be such laws that exploit patterns or systematicities with a high rate of precision. But try predicting what an expert will do by taking stock of the entire network and signal strength of neurons in her brain and you will appreciate just how impossibly strenuous and time-consuming that would be (even if it were theoretically possible). We would find ourselves with a lack of time and competence to explain or

[135] There are some caveats that would stop us from focusing on that epistemic agent (namely, reducibility and regressability - they will be addressed in Chapters 5 and 6 respectively), but they are not defeaters of epistemic agency. Instead they are defeaters for a particular epistemic agent postulate being the most appropriate focus.

[136] And given the intuitive agreement of many thought-experiments in epistemology, it seems that the epistemic stance is, much like the intentional stance, if not innate, then perhaps deeply rooted.

predict her behaviour through this level. But systematicity can come in various degrees of accuracy. Most physical sciences purport to be exact (which is why Dennett talks of the physical stance). Unfortunately, neither the epistemic nor the intentional stance can boast of such exactness. But that doesn't make them lose their great *approximate* value. And what the epistemic stance loses in accuracy (compared to a physical stance), it makes up for in swiftness. It is a lot easier to explain or predict a subject with the epistemic stance than it is to do the same from the vantage point of one of the physical sciences (e.g. a chemistry stance).

But if we wish to bypass the overwhelming complexity from lower-level patterns, then we need an opportunity to do so. That opportunity is macro-systematicity. Systematicity can come in various kinds. It can be exact or, as in the case of the epistemic stance, approximate. While the epistemic stance may not be as precise as the physical stance, we will see how it makes up in swiftness what it loses in precision. But let us unpack the concept of "macro-systematicity" a bit further. "Systematicity" refers to a pattern which a theory can exploit and "macro" refers to the larger or higher level (relative to another level) where we find this pattern. To make this more intuitive, consider the following drawing:



(Hofdstadter, 1999, p. 310)

At the lowest level (discernable here), we can find a pattern, namely the words of "reductionism" on the left, and "holism" on the right. But if we look at the same picture from a higher level, we can see that on the left, these patterns of "reductionism" spell out a different pattern, namely the words "holism". On the right, the patterns of "holism" spell out "reductionism". Furthermore, if we go a level higher still, we can find another pattern, namely the letters "MU". The epistemic stance clearly aims

at a higher level of explanation, because it detect patterns or systematicities that are not atom or neuron-level patterns, even if it supervenes on them.[137]

A stance that reveals such a systematicity at a higher level can be called "a macroscope". The term can be traced to de Rosnay (1979), who used it to refer to a symbolic instrument to see, not that which is small (for which there is a microscope) or that which is far away (for which there is a telescope), but that which is complex. For Rosnay, "macro" refers to what is "too great, too slow, and too complex for our eyes (human society, for example, is a gigantic organism that is totally invisible to us)" (de Rosnay, 1979, introduction), but I'll use "macro" as a higher level of explanation, relative to another level. Any theory that presents an opportunity to bypass micro-complexity could be called a macroscope.[138] The epistemic stance falls neatly under that label, because it bypasses the systematicities at the atom or neuron level to explain or predict larger entities which include those atoms and neurons without focusing on them.

As I will explain in the next section, the explanatory power of the epistemic agent postulate is holistic (as opposed to atomistic) because its explanatory power is spread across acts. But ascriptions are holistic in another sense as well: they are holistic (as opposed to localist) because they spread across their implementation. If the ascriptions were localist, then we could locate each component of the theory in a component of the entity. However, property reductionism must stop somewhere. Brains may store beliefs, but it is unlikely that neurons do and certain that atoms don't. Assuming atoms have beliefs because they make up something that displays beliefs is to be lured in by the *fallacy of division*. The fallacy of division is the error in reasoning where one assumes that what is true of a whole must be transferable to its parts.[139] (Hansen, 2019) Just because a house is large does not entail that its bricks are also large. At some point, while decomposing entities into parts, we lose certain concepts because their implementation is spread or distributed *across* the entity we are decomposing. The largeness of the house is spread across its bricks, and the beliefs of a brain (or mind) are spread across

[137] Supervenience means that there can be no change at the higher level without a change at the lower level. So if the mind supervenes on the brain, then there can be no change of mind without a change in the brain (and possibly other body parts). More on supervenience in Chapter 5.

[138] Theiner (2017) also used the word "macroscope" to consider different forms of cognitive-level systematicities, namely: (i) the intentional stance, (ii) the information processing stance, (iii) the computational stance, (iv) the ecological stance and (v) the dynamical stance. Each of these could be seen as an explanatory approach to epistemic subjects and each supplies their own explanatory focus (most of them being about how the reasoning gets done) and their own explanatory benefits. Nevertheless, the epistemic (or intentional) stance is the most ideal for the purpose of attributing understanding, as it is the one that is the most suited and most intuitive for demarcating a coherent and persisting entity.

[139] The converse of this is the *fallacy of composition*. It falsely claims that just because a part has a property, a whole which is composed of this part must also have that property. For example, just because a house is made of large bricks doesn't make it a large house. Just because atoms aren't alive, doesn't mean nothing composed of atoms can be alive.

its neurons. So the particular attributions we interpret the epistemic agent to have (or the abilities we detect it to display) don't need to be located within physical subcomponents of that entity.[140] As the fallacy of division (& composition) showed, there are different levels of explanation, and each of these different levels can have their own patterns, can have their own systematicities. The success of the epistemic stance in predicting or explaining individuals entails that there is some systematicity which the stance can exploit, but they are spread across their implementation.

Furthermore, because these systematicities are macro (i.e. of a higher level), we don't (necessarily) need to know any of the lower-level patterns or systematicities that implement them to be able to talk about them. The epistemic stance postulates a useful abstraction (i.e. the agent) with no dictates on how this must be implemented physically. This entails that we can exploit such a pattern to explain or predict without needing to know any of the lower-level patterns that implement it. Explanatory success can be *multiply realised*. The *multiple realisability thesis* says that system properties can be instantiated by different kinds of physical structures. Multiple realisability was introduced by Hilary Putnam as an argument against brain state characterisations of mental properties, but they are now commonly associated with functionalist theories of the mind (e.g. Block & Fodor, 1972) because functionalists claim that a mental state is characterised by the way it functions in the system, not the way it is materially constituted by the system. (Theiner & O'Connor, 2010) For example: under a brain-state characterisation, pain could only be shared by humans, various animals or aliens if and only if they shared the same physical make-up and state (e.g. the firing of c-fibres). But what is relevant about pain is not its physical make-up. Even aliens with radically different biochemistry could warrant the pain attribution. (Bickle, 2019) Similarly, the same type of mental properties could be instantiated by different kinds of physical states. And the same software can run on different types of hardware. For instance: a calculator could make use of water currents rather than electrical ones if it were to produce all the same results. A water-calculator will be much slower in practice, but they are both implementations of the *same* calculation. Algorithms are neutral with regard to the substrate that realises them. In other words, the functional model of a computation's capacity can be structurally or causally realised in different ways. This is what is called the multiple realisability thesis (Milkowski, 2013) and it is widely accepted. (Kim, 1992) Multiple realisability leaves open the possibility that epistemic agency might (at least in principle) be realised through several different constitutions or structures (or even social fabric). But for that, they do need to first display the appropriate systematicity. So that is what I will discuss next.

[140] This is because its parts may contribute to the ability in different ways. I will expand on this in Chapter 5 when we discuss the reducibility problem.

Systematicity & Virtual Coherence and Physical Cohesion

So much for "macro." Now it's time to take a closer look at "systematicity." I use the word "systematicity" to mean any kind of direct or indirect pattern[141] which a theory can exploit. Because the epistemic stance has explanatory power, what is being exploited is a systematicity.

As I said in the previous subsection, the epistemic agent ascription was holistic (as opposed to localist) because its explanatory power was spread across its implementation. But ascriptions are holistic in another sense as well: because they are spread across its acts. The attributions are holistic in that their success is to be judged by how well the whole package of attributions fare in explaining or predicting the entity's many behaviours. If the epistemic stance were atomistic, then for every component of the stance there would be an appropriate corresponding behaviour (e.g. believing that p corresponding with a particular disposition, such as asserting that p). The story, however, is not quite that simple. Ascriptions depend on each other (e.g. how a belief that-p manifests itself may depend on whether there is a belief that-q), so it is only by looking at how well sets of attributions fare in accounting or predicting groups of behaviours that we can validate those attributions. For instance, particular beliefs (e.g. that the earth is spherical) do not require pairing up with particular dispositions (e.g. asserting that the earth is spherical). It is the package of ascriptions that does the explaining or predicting of the multitude of behaviours (e.g. not charting with direct lines makes sense if there's also an aim for charting the quickest route, along with the belief that a map is a 2d representations of a spherical earth). So the stance does not quite require a "pattern of behaviour" (Pettit, 2014, p. 100), but a systematicity - as evidenced by the continued success of the stance in predicting or explaining the entity's behaviour. The success of the epistemic stance is shown by the whole behaviour (which makes it holistic), and not (necessarily) by its particulars (which would make it atomistic).

Furthermore, the epistemic stance is a powerful tool to detect salient patterns that warrant considering certain things in the world as separate entities. The components (beliefs, aims, tactics) postulated to explain or predict will tie together one epistemic agent and do so in a coherent way. Allow me to elaborate on each of these separately.

First of all, the virtual components (i.e., beliefs, aims and tactics) that we ascribe to an entity to explain it will have to be relatively *coherent* to be explanatory. The belief in ghosts along with the disavowal of the supernatural cannot, together, explain or predict the behaviour of a subject, because both

[141] This is a reason why I refer to "systematicity" instead of a "pattern," because what is being exploited can fall under a broader scope than what is usually meant by "pattern".

beliefs would lead to opposing explanations or predictions. The epistemic stance is thus systematic in the acts it targets, making sure that there is relative[142] coherence among them. There are no formal criteria to do so, because there is also no direct link between the acts or between the postulated features. A belief in ghosts and the belief that there is a meeting in the faculty at 9 o'clock may have no direct connection, except that they, together, explain why Cara avoids a certain route on her way to the faculty. They are linked through their combined explanatory or predictive success. So even though the epistemic stance's systematicity is not explicit, it does have an inherent adherence to coherence in the acts because, without it, it would not be successful.

Second of all, the epistemic stance links the virtual components of an epistemic agent in a functional way. What ties together an epistemic agent is a set of explanatory components which, as a whole, predict or explain things in the world. For instance: Ororo's belief about the current atmospheric pressure informs her tactics in predicting the weather. If she has a belief about broken equipment, this will inform her beliefs about its readings as well as her aims to repair it, etc. This ensures there is a functional link between the components of the epistemic agents.

This also explains why we don't need a coherence requirement for understanding. As a reminder, the coherence requirement is a proposed condition for understanding (endorsed, for instance, by Kvanvig, 2003 and Ylikoski, 2009) to ensure qualitative understanding or keep out false understanding. I already argued (in 3.3.viii) that not only is the degree of coherence proportionate to the quality of understanding, but also (in 3.3.v) that large degrees of incoherence may result in several epistemic stances being more explanatory (e.g. Henrietta and System Hyde) than a single one (e.g. just Henrietta). This last point can now be seen more clearly through the epistemic stance. It is more explanatory or predictive to consider Henrietta and System Hyde through a separate epistemic stance than it is to see them through one. This would not be true for smaller (and less systematic) forms of incoherence. Furthermore, not all forms of cognitive dissonance will reveal two subjects instead of one. They may merely need some contextual quantifiers for a single epistemic agent. But the epistemic stance adheres to a degree of coherence for its explanatory and predictive powers, even if it doesn't have a coherence condition.

[142] I say "relative" because the requirements of coherence are not consistency requirements as seen in mathematical logic. This seems fair, as our most prototypical candidate for epistemic subjects, human individuals, aren't perfectly consistent or coherent either, entailing that a stringent coherence requirement wouldn't even be passed by human individuals. Here, small amounts of incoherence are allowed to the extent that the incoherent epistemic agent remains sufficiently explanatory as one epistemic agent. (See also next footnote).

So much for virtual coherence. Now, what about *physical cohesion*? Once we have a useful target of the epistemic stance, namely one where the stance gives us explanatory or predictive power, we can look at whatever realises the epistemic agent to demarcate the entity, physically. Because the epistemic stance makes no dictates on how to implement that agent (it doesn't matter what material it is made of, where the components are, what type of things they do separately,..), there are no expectations about how the entity should be built up. Nonetheless, there will only be explanatory power in employing a single epistemic stance if there is physical cohesion. As we have seen before, cohesion is not physical glue holding parts together, but a functional interaction between parts. To resurrect our earlier example: if part of one's brain is outside of one's body but connected up via WiFi, then its functional link remains intact. Therefore, the epistemic stance would be equally successful towards the spread out entity as it would have been for the one where everything was contained inside the skull. For something to be considered as part of the subject, it needs to play a role in achieving its acts. To do so, physical components need to interact with one another. Only if parts of the world interact with one another (where relevant) would it make sense that the realising base for the target of the epistemic stance involves all those parts. Without any physical interaction, we may as well employ two epistemic stances, one to each part.

Therefore, the epistemic stance adheres to all of the requirements that were set up at the start of this chapter. It makes sure the targeted entity coheres in its attributes (otherwise it would generate contradicting explanations), "coheses" in its realising base (otherwise it would make more sense to explain them separately) and persists over time (otherwise there would be no need for explanations). Furthermore, it allows us to conceptualise the inner workings (e.g. beliefs, aims, tactics) from an act-based perspective, and demarcate the agent on the basis of all this (by looking for the realising base for the postulated epistemic agent)..

Admittedly, finding a suitable target works best by approaching it from both directions: you can start from a set of acts or start from a persisting physical entity and you'll have to redraw the boundaries until it becomes one explanatory whole - neither the coherence of the acts or the cohesiveness of its parts can guarantee an epistemic agent by themselves. Two acts may cohere with one another, without being informed by one another if they belong to different entities (Sarah's endorsement of the existence of ghosts does not inform Cara's avoidance of creepy mansions, even though both acts can be considered as cohering with one another). Likewise, two parts may interact with one another, but not cohere with one another because they belong to different virtual entities (e.g. one software program may endorse p, where another software run on the same computer endorses not p). Only if

the entity is both coherent and cohesive will the epistemic stance be successful, and the entity be an epistemic agent.

The epistemic stance was developed largely with human individuals in mind, but its explanatory scope isn't inherently limited to human individuals only. Because systematicity is all that matters, and this systematicity can be multiply realised, it is conceptually possible that epistemic agency might (at least in principle) be realised through several different constitutions or structures (or even social fabric). In other words, organic brains bound by skin and skull do not (at least in principle) necessarily need to be the underpinnings of epistemic agency. We may ask ourselves which other entities (outside of those run by organic brains from within the skin or skull boundary) could warrant epistemic agency attributions such that it would be worthwhile to gauge the quality of their abilities (and thus understanding). Of course, we may find that anything that is not a human individual is principally unable to act in a way such that the epistemic stance would be useful. But if that is true, then we are not in danger of misattributing understanding to unconventional subjects (because they will never act appropriately enough), merely in danger of wrongly expecting them to be worthwhile candidates (i.e. that they would, if they acted appropriately, warrant an understanding attribution).

Armed with the epistemic stance, we can consider why entities beyond human individuals, such as coupled systems, groups and artificial systems, may or may not warrant understanding attributions. We will see that non-conventional subjects bring up further caveats. I will address each caveat in the type of subject where they are particularly relevant, namely: extended understanding and the boundary problem, collective understanding and the reducibility problem, and artificial understanding and the origination problem. In each of these, we seemingly have reason to withhold a change of subject - those reasons being: the possibility to embed the understanding to an individual subject standing in relation to her environment, the possibility to reduce the understanding of a group to those of its members, and the possibility to trace the understanding of an artificial system to that of its author (a programmer). In the remainder of this dissertation, I shall address when and where these reasons for withholding the change succeed and when and where they fail, thus further fleshing out our conceptualisation of a subject with understanding. I begin with extended understanding, because it is directly tied to the demarcation of epistemic subjecthood.

4.4 Extended Understanding & The Boundary Problem

So let us consider the boundaries of epistemic subjects. With the epistemic stance in our arsenal, we now have the opportunity to consider the abilities of epistemic subjects beyond human individuals.

And the easiest place to start would be where we find abilities that are realised by human individuals making use of things in their environment.

Abilities Beyond Individuals

Many of the abilities we today find in science and everyday life are more and more frequently including things from our environment to be achieved. This can range from using pen and paper in order to work out a proof, using a calendar to remember which step in an elaborate experiment one needs to take, using a calculator app on a smartphone to work out how much formaldehyde one needs, using a computer model to predict socio-economic effects[143], using an interactive theorem prover (e.g. Coq proof assistant) to discover new mathematical proof, to using meteorology data collecting and analysing equipment to predict the weather. Of course, not all things in the environment help with abilities for everyone. Someone who uses a calculator incorrectly might make more mistakes than without it. And if the calculator is broken, even its appropriate use will result in inappropriate results. But nevertheless, there are countless examples of epistemic abilities in scientific practices and everyday life that lean heavily on external resources in the environment.

Furthermore, the abilities achieved by the human individual along with part of the environment may be something which neither could accomplish in isolation. For example: A sociologist with dyscalculia may be able to output correlational statistics. She accomplishes this by using Matlab, because without it she would struggle with some of the mathematics. But Matlab struggles with everything else, so it requires the sociologists. Furthermore, by using Matlab, certain patterns, problems and fruitful avenues of future research may stand out to the sociologists that wouldn't otherwise, which leads her to further research as well as further use of the program - and the cycle continues. In short, the results of Matlab could be as guided by the sociologist as the sociologist is guided by the results of Matlab. Now, in some single instances, it may be possible to divide the credit in the same way the workload was divided. For instance, in achieving ability A,B,C & D, the sociologist took care of tasks A, B, C, and Matlab performed task D. But there is no guarantee that we may be able to decompose their combined ability into such neatly delineated components. The interaction between the sociologist and Matlab may become just that: an interaction. This entails that if we would like to decompose the output correlational statistics, we won't be able to divide the work into neatly separated work-packages accomplished by the sociologist and Matlab separately. The abilities accomplished by the

[143] Ylikoski (2014) discusses how agent-based computer simulations can help increase (certain kinds of) understanding by improving (among other things) the scope of what-if inferences and increase reliability for an extended subject.

pair won't just exceed their abilities performed individually, but exceed the aggregation of their abilities considered separately.

So abilities may be realised by more than just human individuals. But what do we do with that information? Would it be more prudent to keep the focus on the individual and change the understanding attribution, or would it be more appropriate to keep the focus on the abilities and change the focus on the individual to a larger subject? If we choose the former and keep our focus only on individuals in isolation, we are presupposing that the brain or skull is the de facto appropriate boundary of a subject with understanding - but seeing as this is exactly what is the open question here, it seems at best premature. If we choose the latter and change our focus to the larger subject, we need a way to make sure that we are targeting a coherent and persisting entity. Here we either need a way to address the abilities while keeping the focus on the individual or we need a way to extend our targets of understanding to include its resource use. In either case, we need to take seriously the role of the environment.

To take the role of the environment seriously, there are two possible routes: consider the subject (and its abilities) as embedded or as extended. In both approaches, we need to take seriously that the realisation of abilities (or cognition or the mind) are not fully isolated inside the head of individuals. In the embedded approach, we do this by taking seriously the role of the environment required for the individual to display her abilities, cognition or mind (an approach championed by Putnam, 1975 and Burge, 1979 for instance). In the extended approach, one needs to take seriously that the role of the environment may not always be significantly different to the role of the human individual to warrant the traditional dividing border between the two (an approach championed by Clark & Chalmers, 1998 and Hutchins, 1995). So, do we embed the abilities of the individual or do we extend the subject? To answer that, it is worth presenting a brief summary of the extended cognition and mind theses, because they have been seminal in opening up the possibility of wider epistemic agents.

Extending Cognition and the Mind

In their paper on the extended mind, Clark and Chalmers (1998) challenged the idea that skin and skull are the appropriate boundary of cognition and the mind. They started their paper with a thought-experiment in problem-solving. Each of three people are asked whether various two-dimensional geometric shapes would fit into depicted “sockets.” They each assess the fit, but arrive at their conclusions via different routes:

(1) Person A assesses the fit by mentally rotating the shape to align them with the socket.
(2) Person B assesses the fit by pressing the rotate button on her computer which can perform and display the rotation.
(3) Person C (as part of some cyberpunk future) assesses the fit by using a neural implant which can perform the rotation.

In each of the three cases, it would be fair to say that the answer was arrived at by a cognitive process (since it performed a cognitive function). We could even go so far as to say the *same* cognitive process was performed (since they have the same computational structure). The only difference seems to be *where* cognition was performed. The role of the environment has long been acknowledged. Under externalism, the role of the environment is shown to be relevant to interpret the subject's cognitive process.[144] But what if part of the actual *cognitive process* is external to the human individual? Clark & Chalmers propose to mark that active role of the environment with *active externalism.* In two of the three cases presented, the human individual in question is linked with an external resource in a two-way interaction to execute a cognitive process. This, according to Clark & Chalmers, creates a coupled system that can be seen as a cognitive system in its own right.

> "All the components in the system play an active causal role, and they jointly govern behaviour in the same sort of way that cognition usually does. If we remove the external component the system's behavioural competence will drop, just as it would if we removed part of its brain. Our thesis is that this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head." (Clark & Chalmers, 1998, p. 8 - 9)

There are some worries with this interpretation. For instance, if the external resource is not portable (and can therefore not easily be coupled), the cognitive processes may come apart too easily to viably consider it as a single cognitive system. This may be a worry if we consider Person B and her computer. Clark and Chalmers (1998) concede that there is something to this objection. Nonetheless, this does not change that the process remains equally *cognitive*, so active externalism would not be undermined. The real take-away from the lack of portability, according to them, is that coupled systems should be *reliably* coupled, so that the cognitive processes won't readily come apart. To combat this worry (among others), consider the parity principle:

[144] The most famous example being Putnam's (1975) Twin Earth thought experiment.

> "If, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process." (Clark & Chalmers, 1998, p. 8)

Interestingly, when it comes to demarcating the physical entity, or deciding what is cognitive, the parity principle avoids proposing a mark of the cognitive - a difficult task, for which any candidate is likely to be either too stringent (only humans are cognitive) or too broad (everything is cognitive) - and replaces it with a call for consistency or parity. The parity principle tells us to consider the boundaries of the cognitive system producing the same cognitive abilities with a similar mode of demarcation.[145] And if Person B's consultation of the computer was a process that was done in the head, as it was with Person C, we would readily accept it as cognitive.

Still, even if an external resource is reliably coupled, that doesn't mean we can't continue to see part of the process as "external." A reliable coupling doesn't stop us from explaining the human individual's acts in terms of internal processes and a series of inputs and actions. Nonetheless, to insist on this approach, regardless of the reliable coupling of a system executing a cognitive process, is to make premature assumptions about what the relevant boundary of the subject should be (i.e. the border-targeting approach, targeting the skin and skull in particular), make things needlessly complicated in doing so (by adding extra steps), and betray a lack of consistency to be able to do so (because we wouldn't insist on the extra step if it were all done in the head). So the cases of Person A, B and C are not relevantly distinct.

Next, what can be said for cognitive processes, can also be said for the mind. Consider now the following two cases: Inga hears about an exhibition at the MoMA, which she decides to go visit. She recalls that the museum is on 53rd street, so she makes her way there. Otto, on the other hand, suffers from Alzheimer's disease, and relies on a notebook to structure his life. He takes the notebook with him everywhere he goes and writes down whatever information he wishes to remember. Hearing about the exhibition, he also decides to go visit it. He consults his notebook, which reveals the museum's location to be on 53rd street, so he makes his way there. Otto uses his notebook to guide his behaviour in a relevantly similar way that Inga uses her biological memory. So if we are happy to

[145] It is worth pointing out some of the objections they anticipated in this and further footnotes: For instance, one may insist that cognitive processes must be conscious, and that "external consciousness" doesn't sound plausible. But not every cognitive process is a conscious one. (Clark & Chalmers, 1998) Similarly, not every relevant part or implementation of an epistemic agent is conscious. And we don't put those parts out of play for individuals either.

explain Inga as having the belief of the MoMA being on 53rd street, we should extend that same epistemic courtesy to Otto. Otto has reliable and easy access[146] to the information because he takes the notebook everywhere he goes, and because he wrote it, he automatically endorses it. In all important respects, insisting that Otto does not have the belief until he checks his notebook - the "Otto 2-step" (Clark, 2008, p. 80) - would be the same as insisting that Inga does not have the belief until she checks her long-term memory.[147] They are both an exercise in explanatory redundancy.

Based on this case, Clark & Chalmers (1998) extract some features for extended cognition and the extended mind. Features which Clark (e.g. 2008, 2010) would later go on to see as criteria. So if there is a coupled system that satisfies the criteria below, you have an extended system with an extended mind:

(a) the resource must "be reliably available and typically invoked" (Clark, 2010, p. 46)
(b) the information retrieved must be "more-or-less automatically endorsed (...) [and] not subject to critical scrutiny" (Clark, 2010, p. 46) along similar lines as things being retrieved from biological memory.
(c) the information "should be easily accessible as and when required" (Clark, 2010, p. 46)

The extended cognition and mind theses have now been firmly established in the philosophy of mind.[148] And yet, the implications of these theses had, for the longest time, received remarkably little attention in the field of epistemology. However, thanks to the likes of Tollefsen (206), Pritchard (2010), and Palermos (& Pritchard, 2013) it has made its way into the epistemology literature on *knowledge*.[149] I would like to address its implications on *understanding* in particular. This focus can also be found in

[146] There are some possible objections based on his ease of access. First objection: there is the possibility of him losing the notebook. This does not, however, point to a relevant difference with Inga: What if a surgeon could tamper with Inga's memory? This possibility alone is not a defeater of her having any of the beliefs that could be tampered with. Second objection: there are potential situations where Otto could not use his notebook (e.g. because it is dark, or because he is in the shower). However, this doesn't constitute a relevant difference either: What if it is a possibility that Inga is asleep or intoxicated? This possibility alone is not a defeater of her having any of the beliefs that were masked by her sleep or intoxication. (Clark & Chalmers, 1998)

[147] Another objection, raised by Adams and Aizawa (2001) or Rupert (2004, 2009, 2013) for instance, could be that the two cognitive processes may not be executed in the exact same way. Biological memory and notebook memory are not functionally equivalent up to its finest grains. Nonetheless, says Clark (2010), they both function, at a higher level of abstraction, as memory. And in the case of Inga and Otto, they both enable the same belief-ascription. For more objections and their responses, see Clark & Chalmers (1998) and Clark (2008, 2010).

[148] Spiders are an example of extended cognition in animals (see Japyassú & Laland, 2017), although not quite in the same way implied by the *Eye's Mind* dialogue.

[149] A recent book collecting articles on the topic is called *Extended Epistemology* (Carter et al, 2018).

Kuorikoski (2013) and Ylikoski (2014), although they focus on extended abilities (abilities realised beyond individuals) without conceptualising (extended) epistemic agency. According to Kuorikoski (2011), understanding attributions to extended subjects can be settled as soon as there are abilities:

> "the only epistemically relevant facts of the matter are that the extended system can reliably answer a range of what-if-things-had-been-different questions about the simulated phenomenon, can successfully infer and explain, but the constrained cognitive system (the human) cannot" (p. 21)

But seeing as understanding attributions usually involve a coherent and persisting entity, an epistemic agent, for which the attributions are explanatory, I do believe there is more to consider than whether there are extended abilities. So our relevant question is whether such abilities reveal valid appropriate targets for understanding attributions. And, as we will soon see, they don't always.

Failures of Epistemic Agency

The epistemic stance allows us to draw similar conclusions to those of Clark and Chalmers. We can reinterpret all of the above through the lens of the epistemic stance and see why it is useful to consider certain abilities as those of an extended entity. If the epistemic stance is successful (i.e. explanatory and/or predictive), it will also be warranted. Because the epistemic stance is implementation neutral, it doesn't matter which things in the world contribute to its success (i.e. its macro-systematicities), as long as they do. If a (human) individual using an artefact produces behaviour that would make the epistemic stance useful, then, together, they constitute an epistemic agent.

That said, that abilities are being constituted by more than human individuals is no guarantee that there is a persisting and coherent entity for which attributions of understanding will be explanatory. It is not difficult to think of instances where an individual using an artefact would fail to make an epistemic stance explanatory or predictive. If Otto only infrequently takes the notebook with him or frequently neglects to consult it when he does take it with him, then most of his behaviour will be largely unaffected by the notebook. The epistemic stance targeting both may strike a successful note in the particular instances in which the notebook was used, but it would have no lasting explanatory or predictive power. If a human individual, coupled with a resource, displays appropriate abilities (of sufficient scope, sensitivity, stability, etc), but only for the brief time where the resource is used, then the power of the epistemic stance is as short lived as the coupling. There is no persisting entity. Therefore, attributing abilities (or understanding) to the coupled system would only explain something

at that particular time. But attributions of understanding (usually) also aim to explain or predict something more lasting, namely the abilities of a persisting entity.

This echoes our earlier problem of a lack of cohesion. If there's no sufficient functional coupling, there's no (persisting) coupled entity. Giving Otto a notebook does not automatically result in a larger epistemic agent. Otto has to use it and do so consistently. This explains Clark's (2010) criterion that the resource should be reliably available and routinely used, but from the perspective of the epistemic stance. If the epistemic stance is only erratically useful, it would be more explanatory to just focus on Otto, and include the notebook in the environment whenever necessary, essentially embedding Otto rather than extending him.

Nonetheless, even *with* a functional connection, we may still see the epistemic stance fail to detect an explanatory entity in the coupled system. Consider Lenny. Lenny is looking for the person who killed her wife. To find out who did it, she needs to gather clues and evidence to synthesize into a narrative. To do so, she is often on the move, going wherever the evidence or clues lead her next. Unfortunately, Lenny suffers from short-term memory loss. To compensate, she has worked out a system based on polaroid-pictures and notes in her own handwriting. If she wants to know which motel she is staying in, she checks in the appropriate pocket for the appropriate picture and address. If she starts interacting with a person, she goes through her pictures to see whether she has encountered this person before, and if so, she checks the back of the polaroid to see who they are, how they relate to her and whether she trusts them.[150] Unfortunately, even if she does consult her own pictures and notes, there are still various reasons why the synthesized narratives she gathers in each moment may be erratic and inconsistent. The reasons can range from her misinterpreting her own notes, expanding on the system in a way that won't have the intended effect, to people knowing about her system making attempts to abuse it (e.g. by crossing out information, taking away pictures, or giving her cause to reinterpret her own words). Each of these can result in a lack of coherent behaviour: An important clue she collects in the morning becomes white noise in the evening. An intention she formulates while inside her motel-room, changes interpretation elsewhere. A person she despises one moment, she trusts completely the next. If that is what happens, then the epistemic stance cannot postulate an epistemic agent that will be explanatory or predictive. Even though she dutifully consults her pictures, the behaviour that results from them may not reveal a coherent persisting entity like it did with Otto. The system she devised would need work before it allows her to act with the (relatively) persisting coherence of an epistemic agent.

[150] This example is directly lifted from Christopher Nolan's film *Memento*.

This echoes the earlier problem of a lack of coherence. If there is no coherence among the epistemic properties revealed by the acts of the targeted entity, then there is no evidence of a singular entity. If the epistemic stance does not have any persisting explanatory or predictive power, it would be more explanatorily useful to focus on Lenny and embed her behaviour in the system of notes rather than seeing them as constituting a combined entity.[151] Clark's (2010) proposed criteria did not address coherence. Given that Clark's focus is on cognition and mentality, and not subjecthood, this seems fair, although it may also be a reason why he appears more vulnerable to attacks of bloating (e.g. Rupert, 2004). The epistemic stance offers an easy way out, however. Reliable and functional coupling is not just relevant between physical entities, but also between the virtual components of epistemic agency. It is not explanatory to attribute a belief to an agent that does not (with any relative consistency) "use" that belief. In other words, if one belief does not inform another, even in any and all cases where it would be relevant, then there is no coupled virtual system, because postulating those beliefs together would not be explanatory or predictive. A belief only belongs to an epistemic agent if it is reliably available and used.

Epistemic Agency Beyond Individuals

So the epistemic stance isn't always successful (or isn't always successful enough, outside of a particular space in time) to warrant positing a larger entity rather than an embedded one. But this does not entail that the epistemic stance is *never* successful for extended entities. The artefacts external to an individual can contribute to a successful epistemic stance towards the two if, together, they act with a persistence and relative coherence. To exemplify the conceptual couplings in which this may be the case, I will now construct and discuss 7 different types of cases in the remaining subsections.

(i) Classic Examples

Let us start with the first three. We already started this chapter with two of the three cases, namely Inga and Otto. Inga understands why a theorem is true, meaning she can consistently work out its mathematical proofs, can explain the outline of the proof to a non-expert, show a reductio ad absurdum if the theorem were untrue, etc. Furthermore she can accomplish all these things without any "outside" help. She works out all the problems by herself and then relays them to us. Otto, on the other hand, has problems with his working memory and can't do any of that. But what if Otto could

[151] However if Lenny indeed suffers from long term memory loss, then the epistemic stance will lose explanatory power even if it is focused on Lenny sans environment. So if we are willing to let the short-termed power of the epistemic stance target Lenny in any situation, then, by parity, we should actually be more open to those cases of short-termed abilities (e.g. Otto with infrequent use of a notebook) discussed earlier in this section.

do everything Inga can, provided he is granted pen and paper (which he always carries around with him)? He can prove the same theorems, give the same outlines, show the same reductio ad absurdums - but only if he uses a pen and paper to keep track of what he is doing. Furthermore, he can do the same things as he would if he were in some future cyberpunk society, benefiting from a neural implant. Compare the presented cases (loosely adapted from Clark & Chalmers, 1989):

(1) Inga
(2) Otto + pen and paper
(3) Otto cyborg

How much understanding is present in each of the cases? Well, the abilities displayed are the same, so it should stand to reason that there is an equal amount of understanding. This means that the only difference between them can be in which (if any) entity is an appropriate target for the understanding attribution. If we keep our focus only on Otto sans pen and paper, Inga warrants an attribution while Otto doesn't. But is it fair to keep our focus on Otto sans pen and paper? We cannot simply assume this is the appropriate entity for our considerations as it is precisely the legitimacy of the boundary which is at issue. Barring important differences in their behavioural profile (which we can reasonably presume not to be the case), the epistemic stance should find an equal explanatory opportunity in Inga, Otto + pen and paper and Otto the cyborg. Otto's behaviour reveals an equally coherent and, to the extent that he keeps using his pen and paper, persisting entity as Inga does. The main difference between them is that Inga achieves her abilities from within her skin and skull, whereas Otto's realising base extends into the environment. So, when it comes to the epistemic stance, there is no relevant difference in the behavioural profile of Inga and Otto (+ pen and paper) that would lead us to successfully postulate epistemic agency for one, but not the other.

There is perhaps one main difference between Inga and Otto. Namely that Otto sans his notebook also benefits from an epistemic stance in a way that is easier to make sense of than Inga sans her temporal lobe. This gives us an easier opportunity to consider Otto sans his notebook.[152] It may furthermore be argued (e.g. Rupert, 2009; Milkowski, 2017) that the role of extended cognition can also be accounted for by a story that embeds an individual (e.g. Otto sans his notebook) within a particular environment. This is, of course, true. The extended epistemic agent (e.g. the coupled system of Otto+notebook) offers no explanatory benefits that are over and above those of an embedded

[152] This does raise the question: How many epistemic agents are there *really?* Yet there is a simple answer. Because the postulate of the epistemic stance is an explanatory strategy and not a metaphysical commitment, there are as many as are explanatorily useful.

epistemic agent (e.g. Otto, embedded with his notebook). The problem is that such an approach constitutes a two-step where we may be equally served by one. Furthermore, it would entail giving up on useful macro-systematicities, macro-systematicities we already intuitively detect and exploit. For instance, we say that Otto remembers where the MoMA is and we say Otto can prove. Translating these claims about an extended Otto (that believes the MoMA is at 53rd street, or understands a theorem) into claims about an embedded Otto (that consults his notebook to find out the MoMA is at 53rd street, or that is able to figure out the steps in the proof given what he has written down prior) is possible, but can lose the conceptual focus of what we want to say.[153] We would have to translate all our understanding attributions into a list of abilities and how Otto's embedding in the environment allows him to display them. This is a two-step where one will do.[154] So nothing explanatory is gained from disallowing the focus on the extended entity, but a more efficient explanation is lost by it.

At the same time, because nothing (except efficiency) is lost by focusing on Otto sans notebook, the context can determine which is our most relevant target. It is certainly worthwhile to consider Otto sans notebook in ways that we wouldn't (as readily) for Inga's temporal lobe - e.g. if we want to know how Otto's biological memory works or what helps him most in remembering, our focus should be on Otto sans notebook. But this contextual interest does not support the notion that Otto sans notebook is the only proper target for understanding attributions (or withholding them). If Otto and his notebook display abilities that reveal a persisting and coherent extended entity, it will be fruitful to consider them as such.

(ii) Social Extension & Acting Base Extension

In the previous examples, the extension of Otto with an external resource was with an artefact. But the "external resource" could equally be another person. Someone can be *socially extended* as much as they can be physically. In their original paper, Clark & Chalmers (1998) already foresaw this conceptual possibility.[155] Since then, the idea of social extension has been explored elsewhere (e.g. Tollefsen, 2006; Palermos & Pritchard, 2016).

[153] This line of argument will get extended (pun not intended) in Chapter 5 when we consider collective understanding.

[154] A two-step may provide further complications when it comes to the focus on understanding. Otto may use the notebook, without fully understanding how it works, what its role and what its contributions are. Kuorikoski (2011) makes a similar observation regarding the use of simulations. If a human-computer pair can answer all sorts of what-if questions, then understanding is present. Nonetheless, the "human may not understand the simulation, [even if] the human-computer pair may understand the simulated phenomenon." (p. 21)

[155] "What about socially extended cognition? Could my mental states be partly constituted by the states of other thinkers? We see no reason why not, in principle. In an unusually interdependent couple, it is entirely possible that one partner's beliefs will play the same sort of role for the other as the notebook plays for Otto. What is central is a high degree of trust, reliance, and accessibility." (Clark & Chalmers, 1998, p. 17)

As an example (adapted from Tollefsen, 2006), imagine Olaf regularly hatches schemes to steal the inheritance of the Baudelaire children with his troop of henchpeople.[156] However, Olaf is a hopeless villain and can never keep up with what the current scheme is or which step of his plans they are executing at any given time. So whenever a scheme is hatched, he lays it all out for his Henchperson of Indeterminate Gender, who then dutifully guides Olaf whenever he has forgotten which step of which plan he is executing. The Henchperson does not contribute anything to the plan, but e is always by Olaf's side, and just relays his plans whenever he asks for it. Olaf, in his turn, always trusts em and automatically takes whatever information e gives as true. For nearly every step in his plans, he relies on em to tell him which step that is. Olaf uses his Henchperson in much the same way that Otto uses his notebook or Inga uses her biological memory. The Henchperson functions as Olaf's external memory. Does this mean that Olaf's (epistemic) agency extends into his Henchperson? This is so much as to ask: is the epistemic stance sufficiently predictive or explanatory when targeting Olaf with his Henchperson? Well, thanks to the Henchperson, Olaf is able to execute his schemes and reveal a behavioural profile that results in a cohering set of epistemic properties: his aims to steal the inheritance as well as the specific sub-aims that are purported to accomplish this. His behaviour will reveal he does have beliefs about which step in the process he is in and what he has done before. Furthermore, because the Henchperson rarely, if ever, leaves Olaf's side, and because Olaf typically relies on em to make his next move, this behavioural profile will persist over time. This entails that Olaf's (epistemic) agency extends into the Henchperson. The same analysis would apply if Olaf were conducting laboratory experiments or constructing complex proofs instead of foiling children out of their inheritance. Henceforth, this leaves us with a fourth case:

(4) Olaf + Henchperson

There is one difference between the case of Olaf and Otto that would be worth mentioning. Let me first sketch the situation a bit further. Olaf has more than a single associate. Furthermore, he doesn't want everyone who has a question about his plans to always bother him with it. So whenever someone wants to know what Olaf is up to, they can ask Olaf, or they can ask the Henchperson. E functions a bit like Olaf's secretary. This entails that, unlike what was the case with Otto, it is not only the realising base that gets extended, but the *acting base* as well. One could ask the Henchperson: what is the next step in the plan? And the Henchperson can, at least on some matters, answer on behalf of Olaf (and tell him afterwards, so Olaf knows he was asked). This marks a difference between Olaf and Otto. The

[156] As is the case in *Lemony Snicket's A Series of Unfortunate Events*.

difference is that for Otto only the realising base got extended, whereas for Olaf both the realising base and the acting base get extended. Otto's notebook never acted on behalf of Otto. The behaviour of the notebook was never considered as Otto's behaviour, whereas some of the Henchperson's behaviour (i.e. that performed in eir role as secretary) can be considered as Olaf's. At least to the extent that Olaf acts along. If Olaf were to disavow eir behaviour (e.g. "I didn't say that, you fool, e did!"), he would break the coherence of their behavioural profile as a coupled entity. But if he doesn't, then the Henchperson can extend Olaf's acting base in the same way as e extends Olaf's realising base.

(iii) Extending Who or What?

We have now entered tricky territory, because what if the Henchperson is actually in charge, even though Olaf thinks he is? The Henchperson could be coming up with all the schemes and just tell Olaf they are his. Just because Olaf endorses (some) of the Henchperson's (speech-)acts, doesn't automatically mean they are part of Olaf's epistemic agency. Therefore, we will need a way to mark *which* epistemic agent gets extended and why, as well as why not. Otherwise, the following problem, brought up by Adams & Aizawa (2010) would be difficult to guard against:

> "Question: Why did the pencil think that 2 + 2 = 4?
> Clark's Answer: Because it was coupled to the mathematician"
> (Adams & Aizawa, 2010, p. 67)

There is something clearly wrong about this, so it uncovers a legitimate worry for us. Furthermore, it is the same worry that would plague Olaf's extension into the Henchperson if the Henchperson is coming up with all the schemes. So if there is a reason why Olaf's earlier extension with the Henchperson (or Otto's extension with the notebook) are justified, we need to point to the relevant difference with Olaf's current extension with the scheming Henchperson (or Adams & Aizawa's case of the pencil extending into the mathematician). The snag lies in what gets extended. Every extension involves an extended realising base (usually involving a human individual), but not every extension involves an *extended epistemic agent* (as we will see in cases 6 and 7), so when it does, it is important to ask which agent gets extended, and why. Consider the following case:

(5) OTTOR + Peter

OTTOR[157], an automated theorem prover, suffers from limited working memory. But a workaround has been found. Whenever it needs to remember more information than its memory-system permits, it prints out the excessive information for Peter. Peter's sole task is to keep track of that information and, whenever prompted by OTTOR for the appropriate piece of it, to type it in for OTTOR to use. Furthermore, whenever OTTOR is at work, Peter is there, and OTTOR takes his typing as trustworthy. Without Peter's help, OTTOR quickly fails at anything more complicated. But with Peter's help, proofs can be found, outlines can be constructed and even simple questions about the proofs can be answered.[158] Peter's role for OTTOR is thus a bit like what the pen and paper did for Otto.

What can we say about OTTOR + Peter? As a coupled system, they display appropriate abilities. So there is (at least some limited quality of) understanding present. Furthermore, they may reveal a coherent epistemic agent. And because Peter always helps OTTOR, that epistemic agency will persist as long as Peter comes into work. It seems fair to say that the coupled system understands. According to what we've seen, OTTOR + Peter are an extended understander. But who got extended? It seems fair to say it is not Peter who got extended, but OTTOR. The only role Peter played was as the keeper of some of OTTOR's beliefs (in the same way as Otto's notebook was the keeper of some of Otto's beliefs). The epistemic agency revealed by the coupled system is a lot closer to OTTOR than to Peter. So it seems fair to say that, if either of them gets extended, it would be OTTOR.[159]

The relevant question is this: When can we extend an epistemic agent? And my answer is the following: When there is an opportunity to describe the epistemic agency of the coupled system by merely adding extra epistemic properties to the existing individual epistemic agent. To do so, there needs to be a large overlap between the extended epistemic agent and the existing individual epistemic agent. In other words, having entity (A) extend with (B) requires that subject (A,B) can be achieved by adding extra properties (e.g. beliefs) to subject (A), while otherwise keeping things largely the same. For example, the epistemic agency revealed by Otto and his notebook is the same as Otto-sans-notebook, extended with the beliefs contained in the notebook. Likewise, the epistemic agency revealed by OTTOR and Peter are the same as OTTOR-sans-Peter, extended with the beliefs held onto by Peter. Therefore, OTTOR is the extended understander, not Peter.

[157] Loosely inspired by OTTER, an automated theorem prover designed to prove theorems stated in first-order logic with equality. (McCune, 1990)

[158] For a focus on the problems of artificial understanders in particular, see Chapter 6.

[159] There is also a similarity between Peter and Searle in the Chinese Room (see Section 3.3.v). If any part of the coupled system warrants extension, if at all, it would be the instruction manual, not Searle.

So far, all discussed cases of epistemic agency have offered the opportunity to extend an existing agent. Otto was extended by his notebook, Olaf was extended by including the Henchperson, and OTTOR was extended with Peter. But what if we can't extend either?

(iv) Emergent Subjects from Extended Entities

Even when a physical base gets extended and results in a successful epistemic stance, it is still possible that the epistemic agent it postulates does not correspond to an extension of (any of) its composing epistemic agent(s). Consider the following case: An interactive theorem prover (a program to assist with the development of proofs via human-machine collaboration) called Coqo[160] collaborates with Otto to find proofs. Together, they can find proofs they couldn't (or wouldn't) by themselves. Some of the abilities they display, together, can be pinpointed as a contribution (and therefore an ability) of Coqo (e.g. mechanically checking a proof for validity) or of Otto (e.g. formulating subgoals), but their discovery-process as well as the proofs that result from them are only arrived at because they work together, interactively (i.e. with feedback loops and not merely by distributing the labour). Neither of them could produce these proofs by themselves. Next, let us assume further that, based on the actions of the coupled system, the epistemic stance reveals an epistemic agent – let us call it Coqto – which allows us to explain or predict the coupled system.[161]

(6) Coqo + Otto = Coqto

Even if the epistemic stance is successful in targeting the coupled system, Coqto, there is no guarantee that the epistemic properties we may attribute to Coqto are easily decomposable into the separate contributions of Coqo and Otto. And if they aren't, we cannot achieve the explanatory entity of Coqto by extending either Coqo with the epistemic properties contributed by Otto or vice versa. Even if there is a degree of overlap between Coqto and Otto, or Coqto and Coqo, there is also a relevant degree of difference. There are beliefs that Coqto can be attributed with that explain neither Otto nor Coqo, and

[160] Loosely inspired by Coq, which "provides a formal language to write mathematical definitions, executable algorithms and theorems together with an environment for semi-interactive development of machine-checked proofs". (The Coq Proof Assistant, n.d.)

[161] Now, this may seem like an easy bit of trickery on my part. I merely decide that they would behave in a way that will make an epistemic stance effective. This says nothing about whether people using resources (like pen and paper, or interactive theorem provers, etc) actually do behave with the appropriate macro-systematicity for the epistemic stance to usefully postulate an epistemic agent. Granted, but that is also not the point. If they didn't, there would indeed be no need to extend the subject. But then the problem lies not with the concept of extended subjecthood, but with its applicability. The point here is not that, for instance, people using pen and paper or a notebook must lead to a successful additional epistemic stance. The point is that if an additional epistemic stance is successful, given the use of pen and paper or a notebook, then we have an additional epistemic agent. Whether it ever is or is not, is an empirical question ("Is it?"); at the moment, we are only dealing with the prior conceptual question ("What if it is?").

there are beliefs that explain Otto or Coqo, but not Coqto. To make use of the explanatory power of Coqto, we will need to postulate it as an *additional* entity - albeit one that relies on Coqo and Otto and partially overlaps with them.[162]

Another way of phrasing the point is to say that Coqto is not just an extended entity (because it physically extends beyond either Coqo or Otto[163]), but also an *emergent epistemic agent*. Extended Emergent Agents have been addressed by Varga (2016) as the "Hypothesis of Extended Emergent Cognition (HEEC1)":

> "An extended systemic property P of a system S is an instance of the HEEC if and only if P is an irreducibly emergent cognitive property that is diachronically novel and does not follow from the features of the parts (either taken in isolation or in constellations simpler than S)" (Varga, 2016, p. 18)

In Chapter 5, I will address emergence and the reducibility problem (its converse), but for now, this is as good a formulation as any to distinguish an emergent extended agent from an extended individual agent. And it certainly applies to Coqto. Not all of Coqto's epistemic properties (e.g. certain beliefs about proof-tactics) will follow from those of Coqo or Otto (its parts), because the acts that make them explanatory are only arrived at *together*.

So, when can we change the subject to an emergent epistemic agent? Well, when there is no opportunity to describe the epistemic agency of the coupled system by merely adding extra epistemic properties to the existing individual epistemic agent. To do so, there needs to be a lack of overlap between the emergent epistemic agent and any of the existing individual epistemic agents that constitute it. In other words, if the agency of the coupled system (A,B) is largely different from both agent (A) and agent (B), then we don't *extend* agent (A) or (B) to (A,B) but *change the subject* to the additional agent (A,B) - an emergent extended understander. For example, the epistemic agency revealed by Coqto is too distinct from either Coqo or Otto individually to warrant extending either of them.

---

[162] Here, it's possible that Coqo's acting base is limited to just Otto, meaning Otto always has to answer for Coqto. The realising base, however, is distributed amongst the two of them in a coupled system.

[163] The "extension" of the "extended entity" here refers to the epistemic subject *physically extending* beyond one of the human individuals composing it, as opposed to virtually extending the epistemic agency of the agents composing it. However, if terminology would be too confusing, I'd happily adopt "distributed entity" (more in line with Giere, 2002a) for those cases where the epistemic agency of the coupled system cannot be achieved by extending any of its composing agents, and leave "extended subject" for those cases where one of the composing agents gets extended.

The same can happen with social extensions. Chapter 5 will delve deeper into the possibility of emergent epistemic agents composed of human individuals, but for the sake of completeness, I will nevertheless shortly address a case here. Consider:

(7) Troy + Abed = Troy'n'Abed

Troy and Abed are great friends, and when they are confronted with a task while they are together (which they usually are), they always help each other out. When they were taught mathematics at Greendale, their constant cooperation was considered endearing and therefore encouraged. This entailed that any mathematical abilities they acquired, they acquired working *together*. This means that without each other, they quickly become lost. Some of the abilities they display, together, can be pinpointed as the ability of Abed (e.g. basic arithmetic) or of Troy (e.g. giving definitions for mathematical concepts), but their proof-planning and the (correct) proofs that result from them are only arrived at because they work together, not because they merely distribute the labour. Neither of them could produce these proofs by themselves, and many of the epistemic properties we may attribute to them (while working together) are not easily decomposable into their separate contributions.

Given this, the same assessment can be made for Troy and Abed as was made with Coqo and Otto. If they, together, make for an explanatory epistemic stance, then its postulate - let's call it Troy'n'Abed - can probably not be achieved by extending either of them. Even if there is a degree of overlap between Troy'n'Abed and Troy, or Troy'n'Abed and Abed, there is also a relevant degree of difference. There are beliefs that Troy'n'Abed can be attributed with that explain neither Troy nor Abed, and there are beliefs that explain Troy or Abed, but not Troy'n'Abed.[164] To make use of the explanatory power of Troy'n'Abed, we will need to postulate it as an additional entity - albeit one that relies on Troy and Abed, and partially overlaps with them.

In Sum

Understanding is always predicated on a subject. But that subject with understanding has unfortunately not been considered in the epistemology literature with equal care as the mark of understanding has. Up until recently, it was often assumed that the targets of understanding attributions are (or should) always be human individuals, but there are cases that challenge that

[164] The epistemic agent of the larger entity may, however, influence the epistemic properties of the smaller epistemic agents. For instance: If Troy'n'Abed decide that the first *Alien* film is the best in the Alien franchise, both Troy and Abed may adopt the same opinion because of that. Nonetheless, they don't have to.

assumption. Many of the everyday and scientific abilities are more and more implemented by more than just individual humans (e.g. groups, artificial or coupled systems), and we need a way to conceptualise this with consistency and without an anthropocentric bias. What we need is a mark of epistemic subjecthood that will help us target a relatively persisting target for who the attributions of epistemic properties (e.g. understanding, beliefs, etc) would be explanatory or predictive. This means that the entity needs to be relatively cohesive (physically or functionally connected in producing the epistemic acts) and relatively coherent (singular in its epistemic identity). Furthermore, this mark needs to withhold needless shrinking or bloating beyond what would be explanatory.

As a mark of epistemic subjecthood, I have defended the interpretationist approach, and more particularly the epistemic stance (the intentional stance with an epistemic focus). The epistemic stance is the strategy of interpreting behaviour by treating it as if the entity were governed by beliefs, epistemic aims (i.e. the kind of results that an epistemic practice values), and epistemic tactics (i.e. any serious systematic attempt to get closer to an epistemic result), as well as any other intentions that play a supporting role in the epistemic agency. It is instrumental in that the sole justification for interpreting an entity as an epistemic agent is the explanatory and predictive success of that interpretation. If seeing an entity as governed by epistemic attributes (beliefs, epistemic aims and tactics) has explanatory or predictive power, then, by definition, that entity is an epistemic agent. The epistemic stance allows us to conceptualise the inner workings (e.g. beliefs, aims, tactics) from an act-based perspective, and demarcate the agent on the basis of all this (by looking for the realising base for the postulated epistemic agent).

Epistemic agency can be detected thanks to macro-systematicity. Systematicity is a pattern (i.e. the epistemic agent) that a theory (e.g. the epistemic stance) can predict or explain. The virtual pattern it reveals is not atomistic, in the sense that its components (e.g. belief p) correspond directly to individual acts (e.g. endorsing p), like behaviourism. Instead it is holistic, meaning that components can only predict or explain the behaviour as a whole. Nevertheless, only if the components (i.e., beliefs, aims and tactics) that we ascribe to an entity are relatively coherent, can they be explanatory (e.g. if Otto is ascribed with contradictory beliefs, we will generate contradictory explanations or predictions). The epistemic stance thereby makes sure that the targeted virtual entity is tied together with (relative) coherence.

Once we have a useful target of the epistemic stance, namely one where the stance gives us explanatory or predictive power, we can physically demarcate the entity by looking at whatever

realises the systematicity. Because the systematicity of the epistemic stance is at a higher (macro) level, it makes no dictates on implementation (outside of realising the systematicity). It does not rely on there being a direct correspondence between our attributions of beliefs, aims and tactics and some structure in the brain. The components of the epistemic stance are virtual, not physical. They track something salient in the entity, not because there must be a literal implementation of its structures in that entity, but because it is instrumental in explaining or predicting that entity. Nevertheless, a single epistemic stance will only be explanatory if there is physical cohesion. Only if parts of the world interact with one another would it make sense that the realising base (for the epistemic agent) will involve all those parts (e.g. Otto cannot be ascribed with any beliefs contained in the notebook unless he reads the notebook). Without physical cohesion, it would make more sense to explain those parts separately. The epistemic stance thereby makes sure that the targeted entity is tied together with physical cohesion.

Armed with the epistemic stance, we can consider why entities beyond human individuals, such as coupled systems, groups and artificial systems, may or may not warrant understanding attributions. Many of the abilities in science and everyday life are more and more frequently including things from our environment to achieve them. Furthermore, the abilities achieved may be something which neither the human individual or the environmental resource could accomplish in isolation. Moreover, if those abilities are a result of mutual interaction (with feedback loops), then the abilities accomplished not only exceed their abilities in isolation, but exceed *the aggregation* of their abilities separately (more on that in Chapter 5). To take this role of the environment seriously, there are two possible routes: consider the subject and its abilities as either embedded or as extended. In both approaches, we need to take seriously that abilities are not isolated inside the head of individuals. In the embedded approach, we do this by taking seriously the role of the environment required for the individual to display her abilities, cognition or mind. But one takes it seriously *as environment*. In the extended approach, by contrast, one needs to take seriously that the role of the environment may not always be significantly different to the role of the human individual to warrant the traditional dividing border between the two. I proposed that the choice of approach should depend on the presence and explanatory power of the epistemic stance targeting the coupled system.

Reasons to embed (at least in cases where the environment plays a relevant role) include (i) the resources being so commonplace (for a particular context) that they can be considered as background conditions, (ii) simply being more interested in the human individual as a subject (e.g. to diagnose Otto's memory problems, we need to target Otto sans notebook), (iii) failures of the epistemic stance

due to a lack of cohesion (e.g. Otto doesn't take his notebook with him) or lack of coherence (e.g. Lenny behaves erratically because of the system), or, (iv) the explanatory power of the epistemic stance being outweighed by the explanatory power of another one (e.g. if it takes too long to process). Nonetheless, the epistemic stance can grant us explanatory or predictive postulates (the epistemic agent) that are composed of more than just human individuals. Taking advantage of this power is not only warranted and fruitful, but consistent with our best conceptualisations of individuals. When we do take advantage of this postulate, we are talking about the extended epistemic agent.

Extended understanding means that the realising base of the subject with understanding is larger than the human individual. This always involves an extended realising base, but may involve other extensions, such as the acting base or the epistemic agency from one of its components. Firstly, extending the realising base may also involve extending the acting base (e.g. the Henchperson can let you know whether Olaf is busy or not on a certain day, and both Troy and Abed can speak on behalf of Troy'n'Abed) or it may not (e.g. the notebook does not speak for Otto). In cases where it doesn't, it may seem easy to keep our focus on the human individual, but to do so would be as much of a mistake as to insist one's body is the appropriate target even if the brain that controls it were to be external to it (and therefore, merely part of its relevant environment). Secondly, the epistemic stance taken towards the extended entity may be an extension of the epistemic stance taken towards one of its parts (e.g. Otto gets extended with the notebook, OTTOR with Peter and Olaf with his Henchperson), or may be too distinct from both to consider it as the same entity (e.g. both Troy'n'Abed and Coqto are distinct epistemic agents from the agents that make up). It's time to expand on this last possibility in the next chapter.

PRELUDE 5

A Lovely Forest For a Picnic

Ms. Hare, Prof. Raven and Dr. Ant-Eater[165] *have gathered for a picnic in the forest and Ms. Hare is leading them all to her favourite spot in the forest.*

RAVEN: You won't believe what happened to me though! Yesterday I had an infestation of ants in my office. But the curious thing was that they left an exam-paper on my desk. I don't know which student they got it from, but that student would have gotten top marks.

HARE: Did you consider asking the ants?

RAVEN: I did, but not a single one of them could account for it. Curious, isn't it? Anyway, where's this forest we were going to have a picnic in? All I see is a bunch of trees.

HARE: Yes, that is the forest.

RAVEN: So the forest is just these individual trees? Why didn't you just say that then?

HARE: Because I'm not talking about these individual trees. Any one of these trees may get burned down and replaced and yet it would still be a forest. Sometimes forest fires even do a forest good even if they don't do individual trees any good.

RAVEN: How is it possible that having a forest fire can do a forest any good?

HARE: Speaking of forests, do you know who would have been happy to tag along? Aunt Hillary. She has an excellent eye for forests.

ANT-EATER: Yes, but I'm afraid she's busy with her mathematics exams.

RAVEN: I did not know you had an aunt, Ms. Hare?

HARE: Oh, no, I don't. Aunt Hillary is not really anybody's aunt. But she insists that everybody should call her that, even strangers. She is quite eccentric, but she's one of the best-educated ant colonies I have ever had the good fortune to know.

ANT-EATER: Yes, we have spent many a long evening conversing about mathematics.

RAVEN: I thought ant-eaters were *devourers* of ants, not conversers with them!

ANT-EATER: Oh, I don't converse with ants. They're terrified of me.

RAVEN: Don't backtrack on what you've just said.

ANT-EATER: I'm not.

RAVEN: Then you are being inconsistent.

ANT-EATER: I'm not being inconsistent at all. I said I conversed with Aunt Hillary, but she's no ant, she's an ant colony.

RAVEN: You're just trying to have your Aunt and eat her too. You know there's no such thing as an ant colony. There are worker ants and larvae, and there is the queen. And no ant colony can do anything except through its ants.

ANT-EATER: You could put it that way if you insist on seeing the trees but missing the forest, Professor Raven. But try to think of it through a macroscope.

RAVEN: I see the problem! I've never had one of those. Do you buy them on Amazon?

HARE: You don't buy anything on Amazon if you have a conscience! It is a morally bankrupt company.

RAVEN: That's very unfair of you. I've never met an Amazon worker I didn't like.

[165] The character of the Ant-eater, as well as some lines, have been lifted from a dialogue in (Hofdstadter, 1999) in service of this one.

ANT-EATER: Once again, you're missing the Amazon for its employees. But one metaphor at a time. A macroscope is not a physical instrument, it's a way of seeing. You have to broaden your scope to the macro, the larger. Ant colonies, seen as wholes, are quite well-defined units, with their own qualities - at times including the mastery of language.

RAVEN: I find it hard to imagine hearing an ant colony speak!

ANT-EATER: Don't be silly, Professor. Of course ant colonies only converse in writing. The ants form their words by forming trails. When the trail is completed, I can decode what Aunt Hillary is saying. And if I want to say something in response, I just draw trails in the moist ground and watch her reply taking shape again.

RAVEN: There must be some amazingly smart ants in that colony, I'll say that!

ANT-EATER: I think you are still having some difficulty acknowledging the difference in levels here. You see, all the ants in Aunt Hillary are as dumb as can be. They couldn't converse to save their little thoraxes!

RAVEN: Well then, where does the ability to converse come from? It must reside somewhere inside the colony! How else could Aunt Hillary display any aptitude for mathematics?

HARE: Forgive me for interjecting, but it seems to me that the situation is not unlike the composition of a brain out of neurons. Just like yours, Professor. Certainly no one would insist that your individual neurons have to be intelligent on their own in order to explain the fact that you can be a Professor of Mathematics?

RAVEN: Oh, no, clearly not. With my brain cells, I see your point completely. Only... ants are a horse of another colour. I mean, ants just roam about at their own will, chancing now and then upon a morsel of food. . . . They are free to do what they want to do, and with that freedom, I don't see at all how their behaviour, seen as a whole, can amount to anything coherent - especially something so coherent as the behaviour necessary for conversing.

ANT-EATER: Ah, but you fail to recognize one thing, Raven - the regularity of statistics.

RAVEN: How is that?

ANT-EATER: For example, even though the ant colony can seem like it's just a conglomeration of ants doing their own thing as individuals, there are nevertheless overall trends, involving large numbers of ants, which can emerge from that chaos.

HARE: Oh, I know what you mean. In fact, ant trails are a perfect example of such a phenomenon. You have quite unpredictable motion on the part of any single ant - and yet, the trail itself seems to remain a well defined and stable pattern.

RAVEN: And yet, those trails are made by individual ants. So why aren't the actions that you pretend are those of Aunt Hillary just a shorthand for what the ants are doing? After all, nothing the ant colony does can be achieved without any of the ants actually doing it.

ANT-EATER: Because talking about their team efforts as efforts of individual ants requires that you can explain what the team is doing by talking about what the individual ants are doing. This may work for team efforts that are a mere aggregation of individual ant actions, but there will be certain patterns at the level of the ant colony that you'll never be able to see from just looking at the ants.

HARE: Here we are! We're at my favourite picnic-spot.

RAVEN: Already? But there's nothing about any of these particular trees that is nicer than any of the others we've seen.

HARE: But isn't the forest lovely here?

Chapter 5

COLLECTIVE UNDERSTANDING & THE REDUCIBILITY PROBLEM

In the last chapter, we ended on attributions of understanding in cases that required a social extension. As we shall see in this chapter, there are countless examples in natural language where groups are attributed with understanding (see Boyd, 2019). Furthermore, for many of the sciences, the aims of producing understanding could be attributed to several levels: that of the individual, that of a community of scientists, and that of the scientific enterprise as a whole. (de Regt, 2019) Are these attributions supposed to be merely empty rhetoric, superfluous metaphors, or convenient shorthands, or is there any genuine explanatory power to them? To be able to answer that question, we need to bolster our conceptualisation of "collective understanding" and when it is satisfied. While there is a lot of literature to draw from regarding group abilities, group beliefs or group knowledge (and draw from it, I will), collective (or group) understanding has been remarkably absent from the literature. An exception is (Boyd, 2019), who tackles a propositional or representational form of collective understanding (with group grasping) and is therefore only indirectly relevant to us.[166] Nevertheless, I will briefly compare his approach with mine at the end of this chapter.

While I will not conclusively answer whether candidates of group epistemic agents exist, I will shed light on the conceptual space involved in substantiating such an answer. I will argue that a couple of basic steps need to be traversed for a group to warrant the attribute of collective understanding. First and foremost, there needs to be a group, along with whatever that entails. Secondly, that group needs to, as a group, display abilities (because there is no collective understanding without the trait of understanding). And thirdly, those abilities need to result in a successful epistemic stance (because there is no collective understanding without an entity to attribute it to). However, even if a group of human individuals forms a body that acts as one (thus creating an explanatorily powerful target of the epistemic stance), it may yet be possible to reduce that group-level explanation to individual-level explanations, making the appeal to a collective subject superfluous. When is such reducibility a problem and when is it not? I shall argue that reducibility is a problem when the abilities and epistemic agency of the group can be straightforwardly mapped onto a conglomerate of those of its members. So the last step will involve the lack of such a mapping relation (because there is no collective understanding if the attribution is not uniquely tied to the group). I will give both an idealised example as well as a brief indication of real world examples. Let us now consider each of these steps separately.

[166] Another exception is, not surprisingly, (Delarivière, 2020) - which can be seen as a blue-print of this chapter.

5.1 Epistemic Group Abilities

The first steps are quite straightforward: For attributions of understanding to be meaningful, the entity in question needs to be a *group* that displays the systematic and valuable *trait* of understanding. In Chapter 1, I argued that the trait of understanding is best characterised as abilities (or multi-track acts). This means that we couldn't speak of collective understanding unless there were abilities. And the only potential candidates for collective understanding are, not surprisingly, groups. So how can groups have abilities? To address this, we first need to clarify what we mean by group and how they can display abilities.

Demarcating Groups & Member Contributions

The only potential candidates for collective understanding are *groups*. So what is a group? A group is constituted by a "set" of individuals which are its members. "Set" should here be read in the most liberal sense of the word and not in the mathematical sense. A set, in the mathematical sense, is composed of a fixed membership, whereas a group can survive a change of membership. (Epstein, 2018; List & Pettit, 2011) For example, if you take a mathematical set such as the set of natural numbers, and remove the number 2, the resulting set no longer has the identity of being the set of natural numbers. However, if you take a group, such as the board of directors of Wayne Enterprises, and remove (or replace) one director, the group keeps its identity as board of directors. So we need a different way to mark what is a group.

There have been different proposals in the social ontology literature on how best to define the identity of a group. (see Epstein, 2017, 2018) Sometimes distinctions are made between groups and other social entities, such as a aggregates (which pick out every entity to whom a label is appropriate, e.g. red-haired women), a collection (which change identity with a change of members - List & Pettit, 2011), a corporation (which are groups that have a structure and decision-making process, e.g. the government - Tollefsen, 2015), an institution (which has institutionalised aims and procedures), an institutional person (which is a group with an aggregation procedure that avoids the discursive dilemma - Pettit 2003), etc. Different kinds of groups will require different demarcation approaches to figure out who does or does not belong to it. So what makes a group, a group is different for "red-haired women" than it is for "Harley's book-club" or "Wayne Enterprises" or "Gotham's Crime Investigation Unit" or "visitors of the MoMA museum" or "CERN" or "the sixth graders." Even within a particular kind of group, many distinctions can be found (some of which we will see in the next section). For our present purposes, I want to cast a wide enough net to capture any type of tying-together of people that would in everyday language be labelled as a "group." This is a low bar, because

there are a number of reasons why we might label a certain number of people as a group. Yet this needn't worry us, because the relevant criteria that will help us narrow down the appropriate targets for collective understanding will be provided by the requirement for epistemic agency and irreducibility of the group agent, not by a demarcation-criteria for the notion of group, which is really only needed to make sure several individuals are involved.

If we want to know whether any group of individuals can be attributed with collective understanding, that group will need to first and foremost display the *trait of understanding*, namely the appropriate abilities. If a group displays no abilities, then an attribution of "collective understanding" would fail, not so much because the group isn't a useful or unique target for our attributions, but because nothing about them would prompt us to attribute understanding in the first place. Even a misattribution couldn't get off the ground, because there is nothing that would make us look for anything to attribute understanding to. But to know whether a group can display any abilities, we need to have a better sense of *how* a group can display any abilities (regardless of whether those abilities are uniquely those of the group or not). I will once again cast a wide net by saying: Acts are of a group if they are carried out or contributed to by its members *as (or within) part of their role as a member*. De Ridder (2018) makes a similar appeal to intuition when he says: "the intuitive idea is that everything that a group does qua group and everything that the group members do qua group members is part of the group's life." (p. 49) And this intuitive idea is what we will be operating with.

Nevertheless, the constraint of "as (or within) part of their role as member" is relevant if we want to make sure that not every act that can be pinpointed in a member becomes an act of the group. Here is a simple example: the group "visitors of the MoMA museum" roughly corresponds to the people being present in the MoMA museum at a particular time (excluding people who are there with a purpose other than visiting, such as guarding, cleaning or curating the museum). Even in this un-excitingly simple example, we can distinguish between a member's behavioural profile as an individual and their behavioural profile within their role as a member of the visitor-group. If their membership is determined by visiting the MoMA museum, then only those acts and contributions made while visiting the museum will be part of their contribution *as a member*. So, for example, we could say that the visitors of the MoMA are disposed to produce drawings (because many aspiring artists can be found making sketches of the works on display), but they are not disposed to predict the weather. This remains true even if most of the people visiting the MoMA turn out to be meteorologists. Unless they predict the weather *while visiting* the MoMA, their tendency to predict the weather as individuals

has nothing to do with their membership of the group "visitor of the MoMA museum".[167] Conversely, acting outside of the membership does not contribute to the acts of the group. For example, all the members of Harley's book-club may be in the same supermarket at the same time, but they are not shopping *as members* of the book-club. So Harley's book-club is not out shopping even though all of its members are.

In the visitor-case, it is fairly straightforward to distinguish between the acts that count within the role of member, and those that don't (namely acting while at the MoMA[168]). In some cases, the demarcation between what counts within the role of membership and what doesn't will be more difficult, vague or complex. Membership of a group may influence behaviour (e.g. being religious) or not (e.g. eye-colour). Furthermore, it must be noted that what counts as "acting within the role of a member" does not require that the members recognise their status as a member, or see their contribution as a member-contribution (e.g. some men who display toxic masculinity). The label of "group" often comes from recognising a pattern or macro-systematicity and then finding out who contributed to that pattern (and is therefore playing a role as member). Depending on the nature and purpose of the group, the role of membership could be an easily demarcated feature (e.g. the red-haired), a past act (e.g. divorcee) or current location (e.g. those inside the museum).[169] Most of the time, however, the defining role membership of a group is trickier (more vague or more complex) than this. What constitutes acting as a member of Harley's book-club, for instance, is not so easily pinpointed. If we say that the membership of Harley's book-club is defined by discussing the book while at Harley's house, then the book-club would disband with every action that is not discussing the book, and convene at any point that any arbitrary person discusses the book in her house. What counts as a role for membership may comprise several actions, decision-procedures, public commitments,

[167] We may of course use the phrase "of the MoMA can predict the weather" in the sense of "there are visitors of the MoMA who, when outside of the MoMA, are able to predict the weather," but then one is pointing to there being an overlap in the set of people whose membership is comprised of being able to predict the weather and those whose members are comprised of visiting or *having visited* the MoMA, rather than as an attribute of the latter. This is similar to phrases such as "visitors of the MoMA tend to be highly educated" or "visitors of the MoMA also visit the Guggenheim museum," but this does not entail that getting a university degree or visiting the Guggenheim museum is done as a member of the MoMA-visitors group. It is merely pointing to the overlap between people who *have* visited the MoMA and people who have X (e.g. acquired a university degree, visited the Guggenheim).

[168] Perhaps it would seem tempting to say that a person is a visitor of the MoMA only in paying the visiting fee and looking at the art-work, but then we are restrictive up to a point that nothing interesting can be discovered or said about the visitors of the MoMA outside of their paying the visiting fee and looking at artworks. Phrases like "visitors of the MoMA museum leave behind a lot of trash" would be impossible, as those acts are not part of the defined acts of visitor. Yet it is intuitively clear that it is within their role as visitors that (some) of them leave behind trash. Consider how odd it would sound to claim that it is not the visitors who leave trash (after all, leaving trash cannot be part of an act of a visitor), but merely that there is an overlap in the set of people whose membership is comprised of those who visit the MoMA and those who leave trash at the MoMA. Contrast this case with the one in the previous footnote.

[169] Some of the examples were taken from (List & Pettit, 2011).

adherence to certain procedures or coherence with past actions, etc. However, as the subsequent arguments for collective understanding do not rely on a demarcation-criteria for each group, I won't go into further detail except where necessary, and rely on our shared intuition until this should prove to become a problem.[170]

There is one kind of group in which we are particularly interested in here, which is *epistemic* groups. A group is epistemic if it engages in epistemic activities or displays any epistemically relevant behaviour or abilities. For obvious reasons, we will be particularly interested in groups displaying epistemic abilities. Multiple examples of this exist. From the abilities of a pub-quiz team to get all the answers right, or the ability of CERN (a large group of specialised teams) to publish experimental results (e.g. providing evidence for the Higgs-Boson) (Knorr Cetina, 1999), to the ability of a crime investigation unit (which involves a process that starts at the report of a crime and leads up to prosecution) to find a provable narrative of what happened and pinpoints the ones responsible for it. (Huebner, 2013) But there are multiple ways in which the roles and acts of members acts can contribute to this.

Assembling Member Contributions

Here is an important, but not obvious point: Members don't necessarily have to display the epistemic ability of the group to contribute in their role as a member. For instance: CERN can produce a scientific paper, but this is not because any one of its members produced it. Members don't even necessarily have to act epistemically to play their role in the group's epistemic act or ability. Those members of CERN who are control operators or responsible for logistics or Human Resources may, in their role as member of CERN, not act in any way we could safely call epistemic, and yet their contribution (e.g. through scheduling, providing instruments for information transference, making sure the accelerator is operating smoothly, etc) may be vital for CERN to act epistemically. That said, members can also play their role more directly. They can act on behalf of the group, be it by virtue of being a member (e.g. "CERN is friendly or rude to outsiders") or with specific intentions to act as a member (e.g. endorsing the group view as a member, even if it is not personally believed) or on direct behalf of the group (e.g. spokesperson for the group as a whole). So even while acting within the confines of membership, there are multiple ways in which contributions can be made to the group. I will distinguish six ways (based on the taxonomy of Steiner and Laughlin discussed in Theiner, 2017) in which members can contribute to group acts: The group act may be carried out by a representative

[170] If all that stands in the way of attributing a certain collection of individuals with "collective understanding" is the applicability of a stringently technical interpretation of "group," then I'd happily forgo the technical term of "group".

member disjunctively, through an additive or conjunctive sum of member contributions, through a compensatory function of member contributions, through a cooperative succession of member contributions or through feedbacked cooperation among members. Each of these deserves a brief elaboration.

In the *disjunctive* case, each group act or ability is contributed by a representative member. This is, in essence, a division of acts or abilities among members. For example, imagine a pub quiz where the members of a team have decided beforehand that all of the culture-questions will be answered by member A, the sports-questions by member B, the history questions by member C and the political questions by member D. They will behave in such a way that for each question, there is a member whose answer will be that of the group. Note that their answers won't necessarily cohere - especially if the members do not communicate with one another (in other words, do not "cohese"). For example, imagine the quiz involves the following question: "When did the prohibition end in New York?" This is a history question, so it will go to member D. For some reason (either because member D doesn't know much about American history or because the division of labour was ill-chosen), member D does not know the answer and guesses 1975. However, if the following question were to be asked: "Name a film with several remakes that came out in the year the prohibition ended" the task would go to member A. Member A knows prohibition ended in 1933, the year King Kong was released. She knows this because in the 2005 remake, the prohibition gets mentioned explicitly. So, strangely, the group is able to name the year prohibition ended depending on whether it is asked a history-question or a culture-question. In a disjunctive case, the group acts are, at best, as good as those of its most capable member (Steiner, 1972) - with the important caveat that this is only true if the disjunction of labour is such that the task is "disjuncted" to the member that is most capable of taking it on.

In the *conjunctive* case, each group act or ability is contributed to through the conjunction of member contributions. For example, the data (as opposed to the papers) generated by CERN is conjunctive. Every piece of data could be generated by an individual member, but the data-pool as a whole is a conjunction of what the members have generated.[171] Here too, there is an important caveat. If any data, generated by a member of CERN, fails to be added to the server for some reason, then their member contribution is (to this extent) not fully conjunctive. In the conjunctive case, the quality of the group acts are constrained by its *least* capable member (and how much that member contributes).

[171] Technically, data generation in CERN will often be a team-effort, but if and where the data is generated by single individuals, they will contribute to the group conjunctively.

Similarly, in the *additive* case, each group act or ability is contributed to through a sum of member contributions. The distinction with the conjunctive case seems to be that the components of what is being contributed is not worth distinguishing. A non-epistemic example is money raised for charity. No distinction is made between the particular money raised by member A with the particular money raised by member B. If data were equally interchangeable, it would also be additive. Every piece of data that would then be generated, for instance by a member of CERN, gets added to the additive pool of data.

In the *compensatory* case, each group act or ability is the outcome of a function of member contributions. For example, imagine a group of people are made to guess the number of balls in a jar, and an average is made of all their answers. If the average of all their answers gives them a good answer, then the ability of the group is due to a compensatory process. (Surowiecki, 2005) But averaging is just one of many possible functions. The group act may be equally formed through unanimity or plurality decisions.

In the *cooperative* case, each group act or ability is the outcome of member-contributions which rely on the contribution of other members in a linear succession. The easiest example here is that of an assembly line. Let's say that the members of a group science-project have decided to divide their labour into successive steps. Member A sets up the experiment, member B runs it, member C interprets the data, member D shapes it into a paper. There may be multiple reasons for doing this. The division of expertise may be purely to lower the workload with a division of labour, or because each member gets assigned with the labour that most closely suits their expertise. What makes this example a cooperative case is that each member contributes to the process, but also relies on the contribution of the members that came before. Member D relies on Member C having done her part of the work before she starts (or can start), whereas Member C relies on the contribution of Member B and so forth.

Lastly, we have the *dynamical* case. This one is similar to the cooperative one in all but the linear succession. Here the outcome is one borne of mutual interaction between members (with varying degrees of complexity in that interactive process). For example, consider the following situation:

> "Suppose we are spending an evening with Rudy and Lulu, a couple married for several years. Lulu is in another room for the moment, and we happen to ask Rudy where they got the wonderful [porcelain] Canadian goose on the mantle. He says, "We were in British

> Columbia .. . ," and then bellows, "Lulu! What was the name of that place where we got the goose?" Lulu returns to the room to say that it was near Kelowna or Penticton-somewhere along Lake Okanogan. Rudy says, "Yes, in that area with all the fruit stands." Lulu finally makes the identification: Peachland. In all of this, the various ideas that Rudy and Lulu exchange lead them through their individual memories. In a process of interactive cueing, they move sequentially toward the retrieval of a memory trace, the existence of which is known to both of them." (Wegner, Giuliano & Hertel, 1985, p. 256-257)

This is what Wegner, Giuliano, and Hertel (1985) termed a "transactive memory system."[172] They define such a system through two essential components:

> "(1) an organized store of knowledge that is contained entirely in the individual memory systems of the group members, and (2) a set of knowledge-relevant transactive processes that occur among group members. Stated more colloquially, we envision transactive memory to be a combination of individual minds and the communication among them." (Wegner Giuliano & Hertel, 1985, p. 256)

Together, Rudy and Lulu are able to tell you where they got that porcelain goose. Their answer was not formed through linear contributions, but through interactive ones, based on knowing (or simply successfully relying on), at least to a degree, what the other one knows.[173] What is most relevant here is that the way in which members of the group contribute to the group's acts is through *mutual interaction*.

These were, broadly, the six different ways of assembling member contributions. Multiple modes of assembly may occur within a single group at any point. The members of a research unit may, for example, generate data conjunctively, trash additively, set up experiments cooperatively and write out papers dynamically. Furthermore, each of these separate tasks (e.g. publishing papers) may have been accomplished by several of these modes of assembly at different (or parallel) stages. The focus of this chapter will not be to identify the precise modes of assembly in different cases, but to explore what these modes of assembly may entail for attributing collective understanding.

[172] See (Palermos, 2016) for a dynamical approach to groups and transactive memory systems.

[173] Some readers may already have picked up that this could be viewed as a case of socially extended memory. For an extended knowledge approach to this example, see (Palermos & Pritchard, 2016)

Before I move on to the consequences of some of these modes of assembling member contributions, I would like to make a brief note on the social features of groups that underpin it. In the social epistemology literature, a lot of attention is given to the social features that could help bring about some of these contribution structures. Examples are: social commitments, common knowledge (of, for example, commitments), shared goals (of, for example, the group's acts), joint intentions, plans for cooperation, etc. While I believe this is an incredibly interesting avenue of research regarding the implementation of social groups, I do not believe any of them are strictly relevant for collective understanding, even if they are useful in knowing how to implement it. It may indeed prove difficult for a group to display any abilities without a shared goal amongst its members, or common knowledge of their commitments, or a plan for cooperation, and this is valuable knowledge in social epistemology. However, as future examples will show, none of these social features are necessary conditions for collective understanding. They are helpful, but not necessary, and certainly not what *defines* collective understanders. Therefore, I have decided to not address them except where they are indirectly relevant.

Assembly Bonus & Loss Effect

In the latter three modes of assembling member contributions (compensatory, cooperative and dynamic), an interesting phenomenon may arise. By virtue of the way in which contributions are assembled, it may be that the group displays an ability that is superior to the ability of its most capable member, and is even superior to the total sum of its member efforts. This phenomenon has been dubbed the *assembly bonus effect* (Collins & Guetzkow, 1964) or a *process gain* (Steiner, 1972). The assembly bonus effect can be found in compensatory, cooperative and dynamical cases. I will now discuss that effect in each of them.

In the compensatory cases, the function performed on its members contributions can lead to acts of the group that are superior to the acts of (any of its) members (or its aggregate). This idea is known as "the wisdom of crowds," made famous by Surowiecki (2005). Surowiecki opened his book with an anecdote of a scientist and a weight-judging contest encountered at an exhibition. Interested in how well the average participant would do, he collected the tickets and, much to his surprise, found that the mean estimate of the all participants[174] (experts and non-experts alike) outperformed any single individual expert. So for some tasks, given the right mode of compensatory assembly, the group can outperform the individual. A compensatory assembly can be especially useful because it preserves a

[174] Which he considered to be representative of the *crowd's* view.

greater diversity of opinion and is more resistant to the dangers of conformism and information cascades. (Theiner, 2017)

In cooperative or dynamical cases, the interaction between members can lead to acts of the group that are superior to the acts of (any of its) members (or its aggregate). In the compensatory and dynamical cases, the interaction between members provides an extra opportunity for the group to outperform its members. This is clearest in the dynamical case. If member contributions are complementary, they can add up to more than the sum of their parts (given that the mode of assembly makes good use of how they complement one another). The transactive memory system from earlier is a good example, because "it is just possible that, without each other, neither Rudy nor Lulu could have produced the item." (Wegner, Giuliano, & Hertel, 1985, p. 257). Only together can they access the memory they were looking for. That is why Wegner et al propose that the memory resides in both individuals as a combined system, a transactive memory structure.[175] Thanks to their interaction, they, together, display relevant abilities that are wider and more sensitive than the sum of their individual abilities. In a cooperative case, something similar can happen, except to a lesser degree. For instance: If Rudy always relies on Lulu's help, Rudy might be able to produce more information than either of them would individually, but they would be constrained by how far Rudy could get based on Lulu's initial contribution (in the example of Rudy and Lulu cited above, they would get as far as "in the area with all the fruit stands").

I believe that an assembly bonus like that of transactive memory systems can be achieved even if it is not about memory-retrieval. Communication between two people about separately stored information can provide information that neither ever had, and neither of them grasps fully by themselves. Consider this pairing: A Star Trek fan and a Star Wars fan are put together in a room.[176] They are asked to work together so as to answer questions about the similarities and differences between the two franchises, even though neither of them knows anything about the other franchise. With enough time and care to answer each question, they can prompt each other on issues so as to construct an answer which is coherent with their own understanding of the respective franchises. They will quickly work out that both are franchises set among the stars, but as they look for ways to describe the two franchises, a particular phrasing that may ring true for one member, may not ring true for another, prompting a revision. If the Star Trek fan proposes to describe their similarity as (hard) sci-fi, the Star Wars fan may object that it is more of a space opera. While correcting each

[175] For a more detailed look at transactive memory systems and the assembly bonus effect, see (Theiner, 2013).

[176] One could adapt this example to be about different scientific fields, but the example as it is presented here will be easier to understand.

other's phrasing to construct their combined answer, they can now point to a similarity and a difference, even though neither of them knows or understands how they apply to the other member's franchise. Crucially, neither of them would have been able to give the answer by themselves, or even have been able to construct half of the answer by themselves because. Thanks to the mutual revisions, the constructed answers are not just a conjunction of two descriptions. Note that it is perfectly conceivable that such a type of interaction may end up generating some mistakes (e.g. when each member agrees with the phrasing of a similarity only because they interpret it in a different way) or obvious omissions (e.g. they may never stumble upon the fact that both franchises focus predominantly on the adventures of a single spaceship of which every fan knows the name, namely the U.S.S. Enterprise and the Millenium Falcon). These are mistakes that are unlikely to occur in human individuals who understand the difference between Star Wars and Star Trek. Nonetheless, the point here is not that they, together, constitute a combined system that displays relevant abilities exactly like that of a human individual, but that they, together, display relevant abilities that are wider and more sensitive than the sum of their individual abilities.

Transactive memory systems were just one example of assembly bonuses. We can find many more examples in a variety of cases, ranging from ship navigation[177] (Hutchins, 1995), and Elizabethian theatre practices[178] (Tribble, 2005), to scientific research (Knorr Cetina, 1999; Giere, 2002b), or crime investigation units (Barber et al, 2006; Huebner, 2013). We will come back to the last two of these in the final section of this chapter.

We could see the assembly bonus effect in the compensatory, cooperative and dynamic modes of assembling member contributions. This is because those are the only tasks where member contributions are not entirely based on individual contributions. In the compensatory mode, the function may contribute to the result, in the cooperative mode, the contribution of a previous member

[177] Hutchins (1995) describes the process of ship navigation (of the USS Palau) as a distributed process. This quote is particularly telling: "The safe arrival of the Palau at anchor was due in large part to the exceptional seamanship of the bridge crew, especially the navigator. But no single individual on the bridge acting alone - neither the captain nor the navigator nor the quartermaster chief supervising the navigation team - could have kept control of the ship and brought it safely to anchor. Many kinds of thinking were required to perform this task. Some of them were happening in parallel, some in coordination with others, some inside the heads of individuals, and some quite clearly both inside and outside the heads of the participants." (Hutchins, 1995, p. 5-6)

[178] Tribble uses a model of distributed cognition like Hutchins (which can focus on expert cognition and not just everyday cognition) to look at how actors in Elizabethan times remembered so many lines from multiple roles at any time without rehearsals. She concluded that many of the difficulties are offloaded into a smartly structured physical environment. Following Hutchins, she quotes: "The task world is constructed in such a way that the socially and conversationally appropriate thing to do given the tools at hand is also the computationally correct thing to do." (Hutchins, 1995, quoted in Tribble, 2005, p. 153) Examples include the use of verse, rhyme, repetition and memorable words to aid recall, the use of information underload to avoid confusing one's lines, computational devices such as plots, etch. Some of these have been used in the dialogue preceding Chapter 3.

may help the next perform better than it would in isolation and in the dynamical mode, this effect is compounded if the contributions can loop between several members. In a purely additive, conjunctive or disjunctive case, one could theoretically be able to straightforwardly predict the act of the group based on one's knowledge of how its members would act. So in the pub quiz example of earlier, if we know how member D would respond to a history question, then we would know how the team as a whole would respond to a history question. [179]

The assembly bonus effect has been criticised on the basis that most groups appear to actually perform worse than its most competent member, or the sum of its members. (Theiner, Allen & Goldstone, 2010) Pavitt (2003) noted that group interaction is not a flawless conduit of information. Comparing an ideal case against the actual results of transactive memory systems, he showed that group interaction is not very efficient, and often doesn't even allow a simple pooling of information. From this he concludes that "group cognition is limited by and cannot transcend individual cognition" (p. 598). [180] Steiner (1972) also had a pessimistic view of group processes, and believed they often failed to meet their full potential either due to a loss of motivation or inefficient coordination. He called both forms of *process losses* (Theiner, 2013). They are examples of the flip side of the assembly bonus effect, namely the *assembly loss effect*. The assembly loss effect (or a process loss) occurs when the group, by virtue of its mode of assembly (i.e. its organisation), accomplishes less than its most capable member or the sum of its member efforts. Boyd (2019) unintentionally gives an interesting example of an interaction that involves an assembly loss effect and is the opposite scenario of the Star Trek / Star Wars fans example:

> "Disagreeing Historians: Two historian colleagues, Celine and Tamika, both specialize in Roman history. However, they disagree about many causes of events in the history of the Roman empire, specifically its demise: while Celine believes that invading (...) [forces] was the primary cause, Tamika believes it was widespread government corruption. While these sets of reasons do not conflict with one another, Celine and Tamika disagree about which explanation is correct. As it turns out, they are, to an extent, both right: the fall of

[179] Shaw & Ashton (1976) have investigated the assembly bonus effect in disjunctively organised groups and are led to believe that even here the effect may be at play. However, I believe they fail to account for the interplay of several modes. Their assessment of the experiment is that "when the task is difficult, the group will spend more time attempting to complete the task and interpersonal stimulation should be greater." (Shaw & Ashton, 1976, p. 471) But interpersonal stimulation is a member-contribution too (even if it does not directly contribute to the end-result), and it is not a disjunctive one. Nonetheless, seeing as the assembly of most group abilities do not fall squarely into one type of mode, this result is an interesting example of how a task that was organised to work disjunctively may still lead to an assembly bonus effect in groups.

[180] For an extensive critique of Pavitt (2003), see (Theiner, 2013)

> Rome was overdetermined, and thus while each cause by itself would have been enough to topple the empire, the full account involves multiple causes." (Boyd, 2019, p. 19)

Although Boyd (2019) phrases everything from the point of view of his representational approach (see 5.4), some of the conclusions he draws can be read through our ability approach

> "although they agree that Rome did, in fact, fall, they cannot agree on which reasons support that answer. As a result, the group is unable to provide a consistent explanation (as its members do not agree what such an explanation should be), is unable to draw relevant conclusions from related information (as its members do not agree on what lessons one can draw from the fall of Rome), and is unable to act as a good source of information (as seeking out information from such a group would likely just result in confusion about what the right answer is)." (p. 21)

If Celine and Tamika were tasked with answering questions *together*, their active disagreement would drag each other down rather than lift each other up. Broadly speaking, criticisms of the assembly bonus effect based on process losses are further examples of how the mode of assembly plays a role in the quality of group acts or abilities - and they therefore validate the assembly effect as a whole, rather than undermine it. Furthermore, there's been empirical evidence in favour of the assembly bonus effect. (See, for instance, Laughlin et al, 2006; Woolley et al, 2010) However, the lesson is still that one needs to take into account both gains and losses. How well groups perform is thus the net result of the assembly bonus minus the assembly loss. (Theiner, Allen & Goldstone, 2010; Theiner 2013)

It seems fairly clear by now that groups can display abilities. In fact, they can display abilities beyond (the aggregate of) the abilities of their members. In the cases where this is so, the answer to the question "What made this group ability possible?" is not a particular member ability or even aggregate of member abilities. The group made it possible *together*, because they only achieve the assembly bonus together.[181] As such, if the displayed abilities (borne of an assembly bonus effect) are relevant

[181] Note that this is not quite the same as Wilson's social manifestation thesis, which occurs when "[i]ndividuals have properties, including psychological properties, that are manifest only when those individuals are part of a group of a certain type." (Wilson, 2004, p. 281) Wilson thus merely requires that individuals behave differently as part of the group than they would if they weren't part of the group. The social manifestation thesis relies on minds (it does not apply if individuals do not have minds - see Wilson, 2004, p. 282) and a difference in how they act separately, whereas the assembly bonus effect does neither. It instead relies on the existence of complementary behaviour, regardless of whether this behaviour is of an entity with a mind, and regardless of whether that behaviour would be different outside of a group.

for understanding attributions, those attributions are most appropriate for the group, together, and not for its members separately, or distributively. Nevertheless, the attribution of understanding in such a scenario is not quite an attribution of collective understanding, yet. What it reveals is the realising base of a particular act or ability. It does not, however, reveal whether it is useful to consider the abilities of the group in a similar way that we consider abilities of human individuals, namely as persisting attributes of a coherent epistemic agent. For that, we move on to epistemic group agents.

5.2 Epistemic Group Agents

Now that we have considered how a group can act and have abilities through the contribution of its members, could these abilities also reveal an *epistemic group agent*? This question is what I will focus on in this section, leaving open whether the supposed epistemic group agency is the best explanation or whether it is reducible to a member-level explanation (which is a matter that we will get to in Section 5.3).

Demarcating Collective Agency

In Chapter 4, we considered what it was that characterised an epistemic agent. As a brief reminder, an epistemic agent is nothing more or less than the successful target of the epistemic stance (i.e. the intentional stance, except with a focus on the epistemically relevant properties). The epistemic stance was the strategy of interpreting the behaviour of an entity by treating it as if it were governed by beliefs, epistemic aims, and epistemic tactics (i.e., any serious systematic attempt to get closer to an epistemic result), as well as any other intentions that play a supporting role in the epistemic agency. For the notion of an *epistemic group agent* to make sense, the group, as a group, would need to act in such a way that postulating beliefs, epistemic aims and epistemic tactics would have explanatory or predictive power.[182] There would need to be explanatory or predictive power by focusing not just on the members and their acts (or contributions), but by, additionally, focusing on the group as a whole. This is an interpretationist approach because instead of starting from a metaphysical theory of cognition or agency, one starts from the success of its ascription. The interpretationist approach has been utilised successfully in social ontology by Tollefsen (2002, 2015) and List & Pettit (2006)[183] and I will continue along that line (although my account will depart from both of them at different stages of the chapter). Tollefsen (2015) says:

[182] There are accounts that focus on group beliefs (e.g. Tuomela, 1992, 2013; Gilbert, 2000, 2004, 2013) without invoking agency. Although these accounts have explanatory value, they focus more on *how* a group believes, rather than *why* it believes. (Tollefsen, 2015) As such, I will touch upon these accounts only where they are relevant to this approach to collective understanding.

[183] And to some extent, Huebner (2013).

> "If we view our practice of making sense of certain groups as agents as an extension of our practice of making sense of others, then the attitudes we regularly ascribe to certain groups are to be identified not with sets of individual attitudes that are interrelated in various ways but with dispositional states of the whole group. Our practice of attributing such dispositional states is guided by norms of rationality and its attendant assumption of a rational point of view. If taking the intentional stance towards groups allows us usefully to understand the group's actions, then we have every reason to believe our assumptions of rationality are justified and that we are dealing with an intentional agent." (Tollefsen, 2015, p. 111)

Because I am focused on collective *understanding*, I have been employing the intentional stance with a focus on its epistemic properties - the epistemic stance. So what would make groups, displaying the appropriate abilities, a successful target of the epistemic stance (which unites their acts as belonging to one epistemic agent)? It is, as I have said, macro-systematicity. Firstly, the "systematicity" refers to a kind of pattern which could successfully be exploited by a theory (even if it is not an academic one), such as the epistemic stance. The systematicity that is required in the epistemic stance is that the postulate of a coherent persisting entity can explain or predict the epistemic acts of the group. The epistemic stance was holistic, because its components (e.g. beliefs) do not match up with single acts (e.g. an endorsement of that belief).[184] Secondly, the "macro" refers to that systematicity being detected at a higher level of abstraction. When we were talking about individuals, macro referred to the personal level, as opposed to the smaller, neural level. Here it was the individual, not the neurons, which need to display systematicity. Now, when we are talking about epistemic group agents, macro refers to the *group-level,* as opposed to the smaller, member-level. Here it is the group, not (just) its members, which need to display systematicity. This is true regardless of whether the member contributions to the epistemic agent are localist (meaning the components of the theory can be paired up with particular members or clusters of members) or holistic (meaning the components of the theory are spread across the members). In short, if we look at the actions of a group (constituted by its member contributions), then it should behave with the appropriate systematicity at this macro (group-)level.

A big take-away of what we saw about epistemic agency was that it is not marked by a particular physical constitution or structure, but by the explanatory success of the epistemic stance. If two very

[184] As a reminder: This is in contrast to its success being atomistic (meaning components of the theory match up with individual acts - so every belief has a corresponding act, etc).

different physically constituted entities share the same explanatory story, then the very same macro-systematicity is multiply realised.[185] So could it be realised by groups? In theory, it could, because as long as the entity provides the appropriate explanatory power, then it is an epistemic agent. This is true regardless of the constitution, structure or social fabric (be it social commitments, coordination, shared knowledge and goals or not) behind that realisation. So let us take a closer look at the possibility of groups benefiting from the epistemic stance.

Collective Abilities without Collective Agency

For the postulate of a collective epistemic agent to be explanatorily meaningful, there needs to be a relatively persisting and coherent single target. To exemplify this point, we will consider a case of a group where we can detect abilities, but where the abilities can't be meaningfully attributed to the group as such an entity. The problem will therefore not lie in the lack of abilities, but the lack of collective epistemic agency (or epistemic group agency, which I will use as a synonym).

First a small side-note. Because I will explore the concept of collective epistemic agency with multiple example cases, I want to make a few remarks about some of the ideal features of these cases. These features have been implicit in many of the examples so far, but it has become more pressing to make them explicit. Firstly, the examples will be conjured up in such a way that their circumstance of evaluation always involves the easiest probe: filling in an exam. This is a narrative feature for conceptual clarification and not a conceptual requirement for collective understanding. CERN, for instance, is definitely an epistemic group, but its organizational structure is dedicated to producing and publishing data, not solving exams. In real life, CERN's publications (among other things) could be taken as a base for evaluating the group's epistemic abilities. The exam is not a necessary (or even paradigmatic) probe for collective understanding, it is just the easiest one, conceptually, to clarify the differences between several ability-displaying groups. Secondly, the cases we will consider will all presume to test for abilities (understanding) in such a way that any comparison between them will mark a difference between the group's abilities, and not between their tests. For this, let us assume they all get the same written exam. Let us furthermore assume that this exam is set up with incredible care, such that many of the salient abilities (appropriate to the object of understanding for the same context of attribution) are tested for, and that it would be difficult to cheat or take kludged shortcuts

[185] The interpretationist approach would be equally effective on both the collective agent as the individual if one replicates the performance of the other. This idea explains why Pettit (2007) claims that "A group of individuals will succeed in becoming a single agent or agency to the extent that the members can coordinate with one another and replicate the performance of an individual agent." (p. 504) Nevertheless, as we will soon see, it is not *necessary* for the interpretationist approach that a group replicates an individual agent.

(e.g. by memorising a few answers to standard questions). This ensures that the results of the exam mark the differences in their abilities and are therefore a good indicator of understanding.

Given these small side-notes, let us start by looking at a case where epistemic agency fails in an obvious way. Consider:

> *Composite Class:* Every day, arbitrarily chosen people (including experts and non-experts) are put in a class and given one exam to fill in. Each in turn adds her/his own answers to a set of questions on the exam.

What makes the Composite Class a "group" is that the arbitrarily chosen people are strung together under one roof, labelled as a "class," and made to contribute to the same exam. Looking at the filled in exam will reveal that the Composite Class does show that abilities can be present in a group (some of the contributing members are experts, after all). Nonetheless, we have strong reasons not to talk about the abilities (let alone the understanding) belonging to the group, except as a shorthand to talk about some of its members (more on shorthands later). The exam filled in by the composite class consists of nothing more than fragmented answers, a literal composite or conjunction of the individual answers - some of which correct, some of which less so. If you ask "Does the Composite Class understand why X?" then the answer will need so many qualifiers about which conjunction of a variety of acts it displays under which circumstances that it won't allow any meaningful shorthanded attributions of understanding (or even beliefs) to the group as a group. The problem, in short, is that there is no stable entity to which attributions can be made.

Another way to diagnose the problem is by saying the exams do not reveal a (single) *epistemic group agent*. Is this diagnosis fair? Well, let us see how the Composite Class fares under the epistemic stance. The answer is: not well. To employ the epistemic stance towards the group as a whole would be very difficult and most unrewarding. The class does not act like any epistemic agent we know, nor will its filled in exam even remotely resemble that of any paradigmatic epistemic agent. The composite class does not display any macro-systematicity in the way we would expect from an epistemic agent. If we look at how the group responds as a whole (on the macro-level), the answers on its exam are too fragmented because its answers are both conjunctive (not revealing a single answer, but many) and erratic (not coherent with one another, and not coherent over time). Seeing the class as a literal sum of *several* epistemic agents can explain the exam and/or predict future exams of those individual members (per day), but trying to see the exam as the output of a single epistemic agent would

inevitably fail because the exam is simply too fragmented to benefit from an extra epistemic stance fixated on the group as a whole.

We now have a good example of a group where the epistemic stance fails even while it displays some abilities. This forces us to put a constraint on attributions of collective understanding outside of the need for abilities, namely that there must be an epistemic group agent (as revealed by the epistemic stance) for which understanding attributions can be explanatory. But the composite class is not the only type of group, so it is by no means proof that epistemic group agency is impossible. So let us consider a case where epistemic agency succeeds.

Collective Epistemic Agents

If we want to give collective understanding the benefit of the doubt, it would be useful to have an example that can serve as the most conceptually powerful case of it. An example that, conceptually, is the clearest and most powerful version of collective understanding. To supply such an example, I would like to draw heavy inspiration from a paradigmatic individual epistemic agent: a human expert (on a scientific topic which we will leave open). Now, I don't wish to make the claim that a human expert (and especially not any particular one) should be the ultimate and singular standard for what characterises an understander, but if I am secure in the claim that a typical human expert is a good example of an understander, then I can also say that any functionally equivalent entity would be an equally powerful example of an understander. After all, both understanding and epistemic agency are marked by a behavioural profile, so it should follow that two operationally equivalent entities which display the same behavioural profile therefore warrant the same attributions.[186] With this (and other reasons that will be explored soon) in mind, I submit the Expert Planet as the most conceptually powerful case of a group epistemic agent with understanding:

> *Expert Planet:* Every citizen on the planet takes on the roles of a single neuron such that the planet is isomorphic to the brain (and relevant body parts) of an expert. The planet is presented with a vast version of the exam and fills it out with its enormous planet-hands.

Some readers may recognise this case as similar to the *Chinese Nation* or the *China Brain* thought-experiment (not to be confused with the *Chinese Room*), except that the object of understanding is

[186] This is a type of parity principle, where the parity comes from considering a behavioural profile of an entity in the same way as the same behavioural profile in a human individual. De Ridder (2018) makes this principle explicit by suggesting a Modified Parity Principle: "(MPP) If, as a group confronts some task, a part of the group's life functions as a state which, were a state in the head of an individual to function similarly to it, we would have no hesitation in recognizing as a mental representation, then that part of the group's life is a collective representation." (p. 49)

not Chinese, but the area of expertise (which we take to be the one covered by the exam) of the expert it is isomorphic to. The Chinese Nation or China Brain thought experiment has been attributed to Ned Block (1978) even though he did not formulate it with neurons[187], as well as Lawrence Davis (according to Dennet, 1978c) even though he did not formulate it with a group of human individuals[188]. I will let Cole (2014) describe the thought-experiment as it is more commonly known today:

> "We can suppose that every Chinese citizen would be given a call-list of phone numbers, and at a preset time on implementation day, designated "input" citizens would initiate the process by calling those on their call-list. When any citizen's phone rang, he or she would then phone those on his or her list, who would in turn contact yet others. No phone message need be exchanged; all that is required is the pattern of calling. The call-lists would be constructed in such a way that the patterns of calls implemented the same patterns of activation that occur between neurons in someone's brain when that person is in a mental state—pain, for example." (Cole, 2014)

While there are a couple of other differences between the Expert Planet and the China Brain (as presented by Cole), none of them are intended to distinguish the two cases in kind. Instead, the differences are small amendments intended to help guide our intuitions more appropriately. For instance, I have changed the nation into a planet so that we are less focused on the people, processes and politics of real nations. Furthermore, I have kept the tasks for the citizens purposefully vague ("each citizen takes on the role of a single neuron"), so as not to imply that sending phones call could exhaust full functional similarity with the work delivered by neurons in brains. I have also added "and the relevant body parts.".This is not just to account for the planet interacting with an oversized exam-paper, or to make it easier to picture the planet as a full-fledged entity, but also to make sure the brain-processes have an output, and aren't relegated to a mere solipsistic process. Furthermore, it allows us to take into account any possible cognitive roles the body might play in individual human cognition (see the debate on embodied cognition, e.g. Wilson & Foglia, 2015), here incorporated by further citizens and/or by the physical design of the planet. To satisfy proponents of embedded and extended cognition, I am also willing to further amend the thought-experiment to include an oversized environment as it won't change anything about my argument, except of course its ludicrousness.

[187] Block (1978) sets up his Chinese Nation thought-experiment to address machine state functionalism, so instead of being isomorphic to neural functions, the Chinese population is made to contribute to a machine-table detailing an individual.

[188] Davis formulated it with a robot. (Dennett, 1978c)

I would like to address that ludicrousness before we move on to assessing the abilities, epistemic agency and reducibility problem in the Expert Planet. It is indeed a highly implausible scenario to expect to occur. So implausible that one may conceivably question why any ramifications of such a thought-experiment should be relevant to us. The reason why is because if there is any kind of group that would constitute a powerful case of epistemic group agency, or group-understanding, at all, it would be a group that, as a whole, operates *exactly,* even in its most fine-grained detail*,* like an individual expert. One of Rupert's (2005, 2011)[189] objections to group cognitive states is that it is very unlikely that there is any fine-grained similarity between the functional profile of individuals and those of any existing groups.[190] But the Expert Planet is set up so that it definitely does. This means that every kind of act, ability or behavioural profile (from which we discern the abilities, beliefs, epistemic intentions and tactics) we discern in the human expert will be equally present in the Expert Planet. The Expert Planet behaves, as a whole, with macro-systematicity - namely with the *same* systematicity as the human expert that its brain (and body) is isomorphic to. Therefore, the epistemic stance can exploit the same (relevant) macro-systematicities and the epistemic stance would have an equal explanatory and/or predictive power as it would have for the individual expert. The beliefs, epistemic intentions and tactics which we use to predict or explain her behaviour will be of equal use in predicting or explaining that of the Expert Planet. And because the epistemic stance makes no dictates on implementation, it doesn't matter to the stance that the entity is comprised of individuals rather than neurons. In short, everything that made the expert fit for employing an epistemic stance can also be found in the Expert Planet and every reason we have to treat the individual expert as an epistemic agent would apply for the Expert Planet. Failing to recognise this may simply betray an unfounded bias against the mere possibility of collective understanding, however ludicrous or unlikely it may be. This makes its ludicrousness a strength. If you are not swayed to attribute collective understanding even in such an extremely unlikely situation as the Expert Planet, then I am not sure why any other collective effort would. It is, of course, possible that no *existing* group will ever resemble the Expert Planet in any relevant way (and we will take a closer look at existing groups in Section 5.4), but the current strength of the Expert Planet is precisely in revealing what would make a convincing case of collective understanding.

[189] Robert Rupert (2005, 2009, 2011) is one of the leading adversaries of the group mind thesis, so we will be dealing with more of his objections in the upcoming sections.

[190] Rupert (2011) points out that "Given the centrality of the role of representations in cognitive science, one might be particularly worried that the functional profile of typical group representations differs drastically from that of individuals' representations." (p. 634)

Nonetheless, it does bear mentioning that one of the arguments surrounding the China Brain thought-experiment is exactly whether it *is* a convincing case of understanding (be it collective or not). Block (1978) has famously argued against functionalism on the basis that it sounds ludicrous to attribute the Chinese nation with mental properties. However, I will say that (i) the scenario's practical implausibility may contaminate what we think about its conceptual appropriateness (i.e. planets or nations tend not to behave even approximately like an individual expert - which is not the same as their being unsuited to be deemed one if they would) and (ii) attributing the Chinese Nation or Expert Planet with these properties is consistent with our best explanation of individuals (functionalism about other minds and understanders) and it is not clear why Block's unease with the idea should count for anything more destructive to functionalism than that it is *surprising* that it entails there could be epistemic group agents.

This may not prove that the Expert Planet is the ideal case of collective understanding, but it does, at the very least, shift the burden of proof to the opposition. If one wants to insist on the Expert Planet being a case without any collective epistemic agency or understanding, one would need to clarify what the relevant difference is and why that difference is the one that matters.[191]

Now that we have a conceptually convincing example, we can compare it to less ludicrous cases to determine their relevant similarities and differences. Consider first the Summative Class:

> *Summative Class*: The answers that the class fills in will be those that are accepted or agreed upon by all or most of its members.

This example is constructed in the spirit of the *Simple Summative Account*[192] (where a group believes that p if and only if all or most of its members believe that p[193] - Gilbert, 2013). The Summative Class

[191] For example: Wray (2001) has argued that groups cannot be attributed with beliefs because beliefs are *involuntary*, and have an associated *feeling*. Therefore, what groups can do is *accept* a view, but it cannot hold a belief. Nevertheless, none of his proposed differences are sufficiently fleshed out such that they can mark out a definite difference between human and group beliefs. If beliefs are a postulate of the intentional stance (even in human individuals), then all Wray has done is marked a difference in feature, not kind. And there is no reason to suppose these differences matter (or are even true). For instance, are all human beliefs involuntary (and why does that matter)? And are all these beliefs accompanied by a feeling (and why does that matter)? Furthermore, his criticism of group belief does not apply to the Expert Planet, so unless the Expert Planet is a single exception, his criticism won't apply to all groups.

[192] The example also, to lesser extent, evokes Tuomela's (1992, 2013) *Joint Acceptance Account* (where a group believes that p if its operative members, in their respective positions of authority, jointly accept p as the view of the group, and this is common knowledge).

[193] The Simple Summative Account has been famously criticised (e.g. Gilbert, 2013) as being insufficient to characterise collective belief even under a procedure that aims to work summatively. For example, imagine every member of the philosophy department believes that animal products are the largest contributor to climate change, but because they

can boast of one thing that the Composite Class can't: it answers with a single voice. But it is important to consider what that voice reveals. The summative account has famously been criticized by List & Pettit (2011), for leading to the discursive dilemma, because "majority voting on interconnected judgments may lead to inconsistent group judgments even when individual judgments are fully consistent." (p. 46) To exemplify this, imagine if the summative class needed to answer the following questions (adapted from List & Pettit, 2011, p. 45-46):

(1) Are global carbon dioxide emissions above the acceptable threshold?
(2) If the emissions are above the acceptable threshold, will the temperature rise?
(3) Will the temperature rise?

Even if the class consists of reasonable and internally coherent experts, the group answer to each of the questions may be not. Imagine that these are the answers of three individual experts making up the class:

	(1) Emissions above acceptable threshold?	(2) If emissions above threshold, will the temperature rise?	(3) Will the temperature rise?
Member A	Yes	Yes	Yes
Member B	Yes	No	No
Member C	No	Yes	No
Majority	Yes	Yes	No

Note that each member is internally consistent with their answers, so given the epistemic stance, you would be able to successfully predict the answer each of them would give to question number three, given their answers to question number one and two. However, if we look at the answers of the group (where each question is filled in according to what is agreed by most members), this internal coherence was not maintained. This instance of a Summative Class would not fare well under the epistemic stance.

fear the response of their colleagues, they express their concern about fossil fuels instead. Then, in spite of the fact that all members believe the same thing, the group will not act accordingly. The Summative Class therefore deals with member-acceptance instead of member-beliefs.

Nonetheless it must be noted that this problem is not *necessarily* the case for all summative classes. For another example, let us only allow in like-minded individuals so that all or most members share the same views regarding the topic at hand. Let us call it the *Like-Minded Summative Class*. Here the majority view of the group will be the same as those of all of its individual views, like-minded as they are (about that particular topic). In this case, the answers of the group will be as singular (as opposed to conjunctive) and (mostly) coherent as those of its like-minded members. And therefore, the epistemic stance will have an equal explanatory power targeting the group as it will targeting most of its members (as pertaining to that topic). One may object that this makes the epistemic stance targeting the group superfluous (as it will result in the same entity as most of its members), but that is a matter we will get to later (in Section 5.3). For now, it suffices that we can show that groups can exist which would result in a successful epistemic stance, and the Like-Minded Summative Class does. Nonetheless, the (Like-Minded) Summative Class is not a very interesting or relevant example, so let us move on.

We can construct an example that is both more interesting and more true to life than the Summative Class. To make sure that a group will act with a coherent voice that is not necessarily the voice of its members, consider a case where the members take care to actively make sure that they act as a single epistemic agent:

> *Jointly Committed Class*: The members of the class jointly commit to answer the exam as one body.

This case is drawn from Gilbert's (2000, 2004, 2013) renowned account of *joint commitment*. She argues that a group can warrant attributions *in addition to* the attributions to its members. Starting from criticism of the summative account, being neither necessary nor sufficient to characterise the features of the group[194], she introduces the notion of a *plural subject*. For a plural subject to come into being, the members of a group need to express a joint commitment to have the group act as a body that "X". "X" could, for example, be "believing that p". So, for Gilbert (2013) a "group belief" entails that the members "are jointly committed to believing that p as a body" (p. 137). This requires them "to emulate, as far as possible, a body that believes p" (p. 140) or "together to constitute - as far as possible - a body that believes that p." (Gilbert, 2000, p. 41). Once committed, the actions of each of

[194] For meaningful collective attributions, the simple summative account is neither necessary (e.g. voting procedure makes the group act with belief p, but members don't need to believe p, just express their commitment – see previous footnote) nor sufficient (e.g. two groups with same members can have different beliefs if they express and discuss different issues – imagine if Harley's book-club had the same members as the Birds of Prey).

the members, as a member, should reflect the group as a whole. This will involve such things as expressing the view of the group in the appropriate contexts and not calling its views or obvious corollaries into question. If any member fails to act in such a way (as a member[195]), they will have violated their commitment and they will be subject to rebuke for it. (Gilbert, 2000, 2004) Thanks to such a joint commitment, they are constituting a plural subject with its own (collective) beliefs that don't have to reflect those of its members.[196]

From the point of view of the epistemic stance, we may say that the group is an *additional target* for the epistemic stance, and a successful one at that. We can attribute the group with beliefs, aims and tactics and this will have explanatory or predictive power, because members will make sure to act *as a single body* (e.g. by adhering to the beliefs of the group), which entails that they will make sure that the actions performed in the name of a group reveal a coherent group entity. Even though "acting as a single body" is not defined, I believe it is an indirect way of saying (or unintentionally aligning with the notion) that the group, as a whole, should display the macro-systematicity that would result in a successful epistemic stance.[197]

Gilbert's account has been applied to scientific practices, which seem to exhibit at least a fair amount of coherence in its core views. (Gilbert, 2000) Furthermore, joint commitment can account for (some) scientific changes as a process, not of changing individual minds, but of changing their collective mind (where a collective mind is made up of what its members are committed to). This would explain why it is difficult to buck the consensus - not because scientists are overly sure of themselves, but because there is a risk of rebuke to bucking the consensus that they have committed to.

But Joint Commitment is not the only way to make sure that a group behaves in a way that makes the epistemic stance useful. The epistemic stance makes no dictates on how its postulate of epistemic agency must be implemented physically. So groups do not necessarily need to replicate detailed brain-processes before they can be regarded as epistemic agents. Nor do they need to follow any particular

[195] One way to act *outside* of one's membership, Gilbert (2000) says, is to qualify one's acts with "personally speaking" but this may make the other members suspicious and maybe see you as an outsider already.

[196] See, for instance, the previous footnote.

[197] Gilbert could be accused of circularity (see e.g. Sheehy, 2002) because acting as "one body" is precisely what the account is supposed to enlighten us on. Under the epistemic stance "as one body" means "displaying the macro-systematicities for which a single epistemic stance would be successful." Gilbert does not have the epistemic stance in mind, and yet the outcome seems to me to be much the same. The "as one body" entails that the members of the class would act such that they, as a class (and as members of the class), would display an entity with coherent beliefs and acts. To do this, members would need to correct one another in such a way that the outcome of their actions (i.e. the exam) will admit of the same macro-systematicity that would pass the epistemic stance. For example: cohering behaviour based on previous endorsements, such that its views would be predictably consistent.

joint intention (List & Pettit, 2011), shared plan (Bratman, 2013), we-mode (Tuomela, 1992, 2013), joint commitment (Gilbert, 2013) or distribution of labour or sub-tasks (Hutchins, 1995; Bird, 2014). It doesn't matter how they bring the appropriate macro-systematicities about, as long as they do. Tollefsen (2015) calls attention to this same point by saying that "we should distinguish between what makes a group an agent and what mechanisms generate the behaviour that justify the attributions of intentionality" (p. 109) Most accounts proposed in social epistemology are focused on the latter, rather than the former. I have here focused more on the former rather than the latter, because it is only to the extent that these proposed mechanisms result in a group that behaves with the appropriate macro-systematicities that these groups warrant epistemic group agency. And several do, even if only in particular cases.

Given what we have seen, contrast the Composite Class with the Expert Planet, the Like-Minded Summative Class or the Jointly Committed Class from the point of view of the epistemic stance. The composite class did not display any macro-systematicity in the way we would expect from an epistemic agent. The exam, taken as a whole, was too fragmented for the epistemic stance to be successful. We might have been able to predict or explain what was going on with individual members and aggregate these explanations to talk about the class, but seeing the member contributions of the class as one whole yielded no explanatory power at all. The Expert Planet, the Jointly Committed Class and even the Like Minded Summative Class, on the other hand, each (for different reasons) behave with macro-systematicities similar to a human individual, a single epistemic agent.

Nevertheless, one thing that we haven't yet considered, however, is the following: group-level systematicities aren't necessarily discernible *only* at the level of the group. Just because we explain the behaviour of the group by explaining it as an epistemic group agent doesn't entail that we should shift the level of explanation from members to the group as if the two are explanatorily distinct. After all, even though the Like-Minded Summative Class is an equally powerful target of the epistemic stance as most or all of its members, the explanatory story of the group is not distinct from the explanatory story of those members.[198] In other words, we can *reduce* the properties of the group to those of its members. So we are faced with one last hurdle for group-level explanations to be valuable: they have to be more than useful, they have to be uniquely useful. In other words, they have to avoid *the reducibility problem*.

[198] Some readers may recall a similar problem plaguing some extended subjects.

5.3 The Reducibility Problem

So it seems that groups can be a successful target for the epistemic stance. However, whether there is a need to change the subject to the group is another matter. A matter which presses us to ask: Is it ever uniquely useful to postulate an additional epistemic agent, namely the group, or does this new explanatory entity merely refer us to (or distract us from) the proper level of explanation: that of its members? This question is what I will be focusing on in this section.

The Shorthand of Reducibility

We will be in a better position to judge the notion of irreducibility once we also have a better idea of what makes the epistemic properties at one level of explanation reducible to those of another. For that, we have to start with why groups could be reducible. That they may be is no controversial suggestion. After all, if groups are realised by their members, what characterises a group or its behaviour is always determined by what its members are doing.

> "In particular, no group agent can form propositional attitudes without these being determined, in one way or another, by certain contributions of its members, and no group agent can act without one or more of its members acting." (List & Pettit, 2011, p. 64)

This is because the group *supervenes* on its members. This entails that there is no possibility of change at the group-level without a corresponding change at the member-level. If a group acts, it does so by virtue of its member-contributions. Even if a group is to be attributed with beliefs, its members need to have acted in such a way that these beliefs have explanatory or predictive power. So every action or property of the group is only achieved if, and only if, its members did their part. This means that for any change we can point to at the level of the group, we would also be able to (at least in principle) point to a corresponding event at the level of the members. But, the objection follows, if we can give a complete causal explanation of the group couched in terms of the systematicity of its members (along with physical structures), then, by Occam's razor, we don't need to commit to a new kind of entity or macro-systematicity. This is, in essence, Rupert's (2011) *simplicity based argument*. Before I address it, let us unpack it first.

There are roughly two implicit components to (or variations of) simplicity based arguments. The postulated group-entity is rejected either (a) because it is explanatorily *superfluous* (Rupert, 2011) or (b) because it is merely a useful shorthand to talk about its members, a *redescription* of the same thing (Geirsson, 2004). So if we can say everything about groups by talking about its members, then the

postulate of the epistemic stance (the collective entity) is either a superfluous waste of time or a mere shorthand. And if it is the latter, then glorifying that shorthand as an independent entity is simply a waste of metaphysics. Both of these (related) issues lead to a reducibility problem and I will address both components in turn.

The first component or variation is about *explanatory superfluity*. Rupert (2011) invokes Occam's razor to say it "speaks against the positing of additional cognitive states[199]" (p. 635) if we can make do without them. If we already have a complete causal explanation, then positing anything further is a waste of time and metaphysics. Now, does this apply to epistemic group agents?

The explanatory superfluity objection would be relevant for Laplacian Demons, who are unphased by the scale and complexity of causal explanations, and who would happily let go of any higher level of explanation, simply because the laws of physics would suffice to explain and predict everything they would ever encounter. But we are not Laplacian Demons. We find it useful to be able to exploit systematicities at a higher level that allow us to bypass the complexity at a lower one. So in that sense it is certainly not superfluous. This retort alone could end the objection, but I will nevertheless take seriously that explanatory convenience is not the only motivator behind our ontological commitments. It is indeed true that we don't *need* to commit to a different entity. Even postulates of individual minds or epistemic agents could be eliminated (the stance that they can, and should be, is called *eliminativism*). However, because the epistemic agency, or cognition, of human individuals supervenes on the brain (and other parts of the body), one could - at least in principle - supply a complete explanation of the entity's behaviour in terms of its brain (and other parts of the body). Doesn't this entail that, by Occam's razor, the epistemic agent of individuals is equally superfluous as those of groups? Rupert (2011) addresses this counter-objection:

> "The most common way to defend psychology against such wholesale eliminativism appeals to distinctive patterns – patterns in intelligent behavior that have no theoretically unified expression or explanation outside of psychology" (Rupert, 2011, p. 363)

So, according to Rupert, cognitive theories supply an explanatory story that addresses a distinctive pattern (or a distinctive systematicity) that neuroscience does not. Note that Rupert does not mean that cognitive theories address happenings that occur *without* neurons (or other body parts) - which

[199] The arguments used by Rupert (2011) are about cognitive theories which postulate a mind, and not the epistemic stance which postulates an epistemic agent - but the arguments discussed here apply to both, so will be used interchangeably.

would be contra the supervenience thesis - but that they address higher level *patterns* (or macro-systematicities) that are not picked up by neuroscience (i.e. by looking at the level of neurons). By saying this, he softens the nature of the simplicity-based argument by allowing further patterns as long as their causal explanatory work is "*distinctive."* (Rupert, 2005) What this means, precisely, is left open.[200] Nonetheless, I do believe there is some merit to this way of motivating the objection, and I will develop it further.

There is one clear difference between cognitive explanations of human individuals and cognitive explanations of groups, namely that the epistemic agency (or cognition) of human individuals addresses a distinct *kind* of macro-systematicity that the epistemic agency (or cognition) of groups does not - relative to its micro-level. It is easier to see why we don't reduce the epistemic properties (e.g. beliefs) of individuals to the epistemic properties of their parts (e.g. neurons): because their parts don't invite being ascribed with such properties. Neurons, synapses and axons are not even viable candidates for epistemic properties. They simply cannot hold, for example, beliefs. But the parts of a group are individuals and individuals *are* appropriate candidates for epistemic properties, such as beliefs. So, whereas it wasn't conceptually possible for an individual to be a successful target of the epistemic stance *and* detect the same (or similar) postulated properties in its neurons, it is conceptually conceivable for a group to be a successful target of the epistemic stance *and* have those epistemic properties be the same (or similar) as those of its members. Group explanations rely on patterns (as revealed by a cognitive theory) or systematicities (as revealed by the epistemic stance) which are not distinctive from the patterns or systematicities of individuals. According to Rupert (2011), this means there is no need to invoke further cognitive (or epistemic) properties for the group:

> "we should eliminate group cognitive states, *given the availability of other cognitivist explanation*s of the relevant data – those invoking the cognitive states of individuals." (Rupert, 2011, p. 636, italics added)

In the next subsection, I will argue why the availability of other cognitive explanations is insufficient (for instance with cases like the Expert Planet[201]), but for now I wish to concede that in a lot of cases,

[200] He briefly addresses it elsewhere as "patterns in intelligent behavior that have no theoretically unified expression or explanation outside of psychology" (Rupert, 2011, p. 363) I will come back to this later.

[201] The argument, in a nutshell, is this: Just because we invoke the same kind of explanation in both levels, doesn't entail that we can systematically redescribe one level in terms of the other, as Rupert implied. Consider the Expert Planet. Is it superfluous to postulate the epistemic properties of the Expert Planet because we can refer to the epistemic properties of its citizens? It may be composed of epistemic agents, but the contributions they bring to the group is exactly the same as those of neurons so it is not clear why they shouldn't be equally "distinctive".

that availability makes a big difference. Consider the Like-Minded Summative Class. Here, Rupert's objection seems incredibly fitting. Because the answers of the class are those agreed upon by most or all members, and all its members were like-minded, the epistemic stance targeting the group doesn't give us a distinct story from the epistemic stance targeting its members. In cases like the Like-Minded Summative Class, the story of the epistemic stance targeting the group could be a mere shorthand for the epistemic stance targeting its members - which brings me to the second component: the possibility of redescribing shorthands.

If we can systematically *redescribe* group-systematicities as member-systematicities, then the group-talk is a mere *shorthand*. We were never actually talking about the group, except as a convenient label to refer to its members. It is easy to think of examples where the attributions we make to the group are such mere shorthands. For instance when we group individuals according to their shared property, such as when they share a purpose (e.g. "Our class has English at 9:30"), an attribute (e.g. "Our family isn't particularly bright") or even just a space (e.g. "Visitors need to leave the premises by 17:00"). In all these claims we could easily swap the group-label (e.g. "class," "this family," "visitors") with "the members of that group," because the group-talk is a mere shorthand to talk about all of its members. Geirsson (2004) calls this objection the *redescription objection*.

> "We (...) say that even though each member of the admissions committee voted against admission, they deliberated (not the committee) and decided that the strengths of the candidate warranted that she be admitted. But often times, instead of us saying that the members of the committee deliberated and concluded that the candidate should or should not be accepted, we conveniently redescribe what goes on by claiming that the committee so decided. There are good practical reasons for us redescribing what went on. For example, we can get a point across more quickly and with fewer words.[202] But our laziness does not have ontological commitments. Us redescribing what went on does not bring minds into existence." (Geirsson, 2004, p. 3-4)

In this objection, the macro-systematicity is rejected as distinctly valuable because it is just a redescription[203] or shorthand of another systematicity or set of systematicities. An extreme example

[202] The convenience of shorthands should not be dismissed, as it seems to be here. That convenience alone is a valid argument against the superfluity argument (depending on how strict or soft one wants to define "superfluous"). Nonetheless, it may be readily conceded that a useful shorthand is not the same as an ontological commitment.

[203] Redescription could go both ways. There might be a systematic way to redescribe a set of micro-systematicities into a macro-systematicity, or there might be a systematic way to redescribe a macro-systematicity into a set of micro-

of reducibility is The Composite Class, because we refer to the whole group knowing full well that this is a mere systematic shorthand for a component-story: the composite of its members. When we, for instance, say "these are the exam-answers of the class," we can replace this phrase with "these exam-answers are the composite of the answers of member A, member B,... member Z" without loss. As such, the Composite Class is a good example of reducibility. However, it was, as we have shown, still a bad example of macro-systematicity. The Like-Minded Summative Class was a good example of macro-systematicity, but it is also reducible. Even though its singular and coherent voice gives the epistemic stance explanatory power, employing an epistemic stance towards the group still has no benefit over employing an epistemic stance towards (the majority of) its members. Except as a shorthand. We can systematically translate any claim about the Like-Minded Summative Class as a claim about the majority of its members. Even though we *can* detect group-level macro-systematicities, they are dismissible as superfluous on the ground that we can easily reduce them to the same type of individual-level systematicities. If the purpose of the epistemic stance is to explain or predict, then being able to detect the same type of properties and make the same type of predictions (with the same or comparable ease) without going one level higher would make the employment of the epistemic stance towards the group redundant, save as a shorthand or abbreviation. The use of such a shorthand is convenient, but its explanatory power is in no way unique. The reducibility problem is, in essence, that the group-level explanation is reducible if it is not uniquely useful (even if it is explanatorily convenient).

What about the Jointly Committed Class? That class also displayed macro-systematicities, which made it useful to employ an additional epistemic stance and postulate an additional subject. But here too, this fact alone does not entail that we need to shift the level of explanation from members to the group as if the two are explanatorily distinct. Unlike the Like-Minded Summative Class, however, the group-systematicities are not equivalent to the member-systematicities. And yet there is also a way to systematically refer to the group's properties by referring to those of its members. We can explain the group's actions by systematically redescribing it through the joint commitment of its members. Each class answer is one which is committed to by its members so that they answer as one body. In other words, we can reduce the properties of the group to those of its members, namely their joint commitment to the class-answer.[204] Postulating a group entity is a useful shorthand (more useful than

systematicities. The latter is a more difficult task than the former if there is multiple realizability, but I will leave that to the side as it is not important for us here.

[204] Gilbert (2013) has taken care to point out that it is not a conjunction of individual commitments, but a "joint" commitment, but the distinction has to do with who has the right to rescind or rebuke, not with the level of explanation.

with the Like-Minded Summative Class), but not a distinct one. There is always a way to translate it back to what it was actually about: a systematic set of member-systematicities.

So if there is a systematic way to redescribe group-level explanations into member-level explanations, then there is nothing explanatorily unique about the group-level explanation. But, as I will argue, the word "systematic" has been operative here. In the next section I will show why there is no guarantee that macro-systematicities always allow such *systematic* redescription and why this matters. In other words, why it is not always a shorthand. The Like-Minded Summative Class is the shortest hand I could plausibly think of for a group with macro-systematicity. But the hand of group-talk isn't always that short, and sizes vary.

The Longhand of Emergence

So far I have been conceding to the simplicity-based argument that group-systematicities are indeed not uniquely explanatory (even if convenient) if they can be systematically redescribed as member-systematicities. In this section I will argue that the epistemic stance employed towards the group *can* be explanatorily unique to the group-level if it exploits a pattern or systematicity that has no equally powerful systematic counterpart at the member-level. If so, that would evade the reducibility problem because it makes the macro-systematicity irreducible and thus *emergent*.

Emergence is a philosophical term that is often invoked, but hard to define. The general characterisation is that "emergent entities (properties or substances) 'arise' out of more fundamental entities and yet are 'novel' or 'irreducible' with respect to them" (O'Connor & Wong, 2015) In slogan form, you could say that "the whole is more than the sum of its parts." What that "more" means, however, is the tricky part. If groups can have "novel" or "irreducible" properties with respect to their members, then it is clear why emergence could be an important reason to change our focus from members to groups. However, it also puts a lot of stress on what we mean by "novel" or "irreducible." If the characteristics of the groups have to be realised by what their members are doing, then there must always some kind of explanation at the lower level that accounts for the macro-systematicities at a higher one - thus making it "reducible" in some way. For authors wielding "emergence" this has created a palpable tension between either denying a reducibility-story (to the extent of making "irreducible" indistinguishable from "magical") or overstressing the importance of finding one (to the extent that any higher-level explanation, no matter how useful or conceptually distinct, would be considered irrelevant if there is also a lower one, no matter how complex). With my own characterisation of emergence I aim to acknowledge that there is always some relation between the

micro and macro, but that the degree of systematic complexity in that relation should make a difference in how we conceive of the macro.

To explain, let us first focus on emergence in individuals. We have established that human individuals are epistemic agents. What an individual does could be explained or predicted by interpreting her to have beliefs, intentions, and rationality. However, since the macro-systematicities are primarily realised by neurons, there must be *some* elaborate way to talk about beliefs at the neuron level. If an individual supervenes on her brain (and other body parts), then any individual story will have a corresponding neuron-story (along with other physical make-up) with which we can redescribe any particular situation. But, of course, it is not the neurons that have those beliefs. Perhaps we may be able to reduce a person's beliefs to a particular belief-area in her brain, but we can't locate properties of wholes in their divided parts, ad infinitum. At some level, the property of the whole will be spread *across* those parts.[205] Furthermore, there is no reason to suppose that patterns or systematicities at the higher-level can be explained by any *systematic grouping* (or function of) lower-level patterns or systematicities. As far as we know, the beliefs of human individuals have no easy mapping-relation to the patterns of its neurons or any selection thereof, making the endeavour of talking about them at the neuron-level as needlessly complicated, unsystematic, and unrewarding as trying to recount the plot of a film using the pixel-level. Macro-systematicities (e.g. beliefs or plot-points) are useful for explanations or predictions, so they are a kind of pattern or systematicity. But they are not a systematic pattern of the micro-level (e.g. they are not neural or pixel patterns). Therefore, there is no way to redescribe the macro-properties as a systematic set of micro-properties. Some properties will only remain conceptually meaningful at a suitable level of explanation. If we wished to redescribe one level into the other, we would have to do it case by case (e.g. this instantiation of the plot-point is implemented in these pixels), and lose the useful macro-systematicity we detected. It is in this sense, and no stronger metaphysical sense, that the epistemic properties of humans are irreducible to neurons or atoms and thus "emergent." That is why I characterise emergence thusly:

> *Emergence*: An entity has emergent properties at a macro-level if those properties have no straightforward mapping-relation to the properties at a level below.

This characterisation focuses on the degree of systematicity in the relationship between two levels of explanation. If this kind of relation-systematicity is high, then we have an easy way to map concepts of the higher-level to those of a lower (namely via that systematic mapping-relation), but if it isn't,

[205] To borrow Mathiesen's (2006) analogy: "a brick is not rectangular because the molecules that compose it are not rectangular." (p. 163) See also Section 4.3 of this dissertation for the fallacy of division and composition.

there is no explanatorily powerful way to talk about the macro-systematicities at the micro-level (because there is no straightforward way to map one onto the other). The difference between an emergent and reducible epistemic agent is thus not an operational or functional difference, but a difference of explanatory benefit. In short, if reducibility was a shorthand, then emergence is a *longhand*, because to redescribe a situation involving a group-systematicity into one involving member-systematicities in the absence of a systematic mapping relation, we have to rely on long winded redescriptions that only apply to particular cases at particular times. By contrast, reducibility (in its strongest sense) always relies on a systematic mapping relation.

> *Reducibility*: An entity has reducible properties at a macro-level if those properties have a straightforward mapping-relation to any properties at a level below.

We don't ascribe plot-points to films because they systematically account for patterns (or systematicities) in pixel-activity, but because they systematically account for the story (a macro-pattern or systematicity). Even though the story is always constituted by pixels, we have no guarantee that the pattern or systematicity of stories will systematically map onto patterns or systematicities of pixels. A lot of plots are definitely predictable, even though the pixels that constitute them aren't. In other words, we have no guarantee that the systematicity of one level can be discerned as a systematic redescription of the other. Likewise, we don't ascribe beliefs because they systematically account for patterns (or systematicities) in neural activity, but because they systematically account for behaviour (as a whole). Even though behaviour is always constituted through neural activity, we have no guarantee that the pattern or systematicity of behaviour will systematically map onto patterns or systematicities of neurons.

Under this reading of emergence, macro-systematicities are novel or distinct in the explanatory sense, but not in the magical sense. Nothing that constitutes the macro-systematicity is left out of a complete micro-explanation. For instance, nothing of the story is left out in fully describing the pixels of a film, and no beliefs are left out in a full description of the individual's neurons (and other body parts). And yet macro-systematicities are patterns or systematicities and therefore useful for explanations or predictions (e.g. plot-points, beliefs). And if they cannot be discerned as a pattern or systematicity at the micro-level (because there's no systematic redescription), then they are novel or *distinct* in the emergent sense because that systematicity is conceptually tied to the macro-level.[206] I believe this is

[206] This entails that, for Laplacian Demons, emergent properties do not supply extra predictive power, although they might fail to see higher level patterns, and literally fail to see the forest for the trees.

what makes the systematicity unique to, or explanatorily distinct (in Rupert's terminology) from, the level of explanation at which it is discerned.[207] If, in a particular group, there were no systematic way to redescribe the group-systematicity into member-systematicities, then that systematicity is conceptually tied to the group-level, and then we need to change the subject from the members to the group if we want to talk about it. So when Geirsson (2004) says that "our laziness does not have ontological commitments," he was being unfairly diminutive by implying that committing to macro-systematicities, no matter how conceptually distinct or practically useful, would be lazy and ontologically sterile.

Emergent Group Agents

Labelling a set of members as "one group" can not only have the advantage of bypassing the cumbersome talk about individual members with the swiftness of talking about all of them at once, but the power of the epistemic stance could be due to a macro-systematicity uniquely tied to the group-level, as opposed to a systematic shorthand for the systematicities at the member-level.

In the most straightforward relation, there is a one-on-one correspondence between the group-property and member-properties. Chunking the members into a group saves some labour, but nothing changes conceptually.[208] In the Expert Planet, however, there is no one-on-one correspondence, nor is there any straightforward way to map those group-concepts on the member-level (any more than there is a way to map individual beliefs to neurons), so the macro-systematicities we detect at the macro-level are conceptually tied to that level. As such, if we want to benefit from the epistemic stance's explanatory power, we can do so only by focusing on the group as a whole. The members simply aren't, in any straightforward way, involved in producing the answers on the exam. The mode of assembly for the citizen's contributions to the Expert Planet are, like the neurons in brains, dynamical. The answers on the exam are not the answers of citizens or neurons that were decided to stand for that of the group disjunctively, they are not a conjunction of citizen or neuron answers. The answers are not formed by an assembly line of neurons in a cooperative manner. Nor are they an average, mean or other function of neuron answers. The answers are formed as the end result of a complex and dynamical interaction. What the beliefs of the members are with regard to the exam-answers doesn't (directly) matter. It may indirectly matter in that they might not perform the role of their respective neuron if their opinions or beliefs were otherwise, but it is in the role of a neuron that

[207] Perhaps this is what Rupert (2011) also meant when he said that distinctive patterns are "patterns in intelligent behavior that have no theoretically unified expression or explanation outside of psychology" (p. 636) Nevertheless, he does not believe group systematicities qualify. One level of explanation involving psychology seems to be enough.
[208] Rather like the relation between machine language and assembly language. (Hofstadter, 1999, p. 290-291)

they contribute to the larger whole, not as an epistemic agent with their beliefs taken into consideration.[209]

Interestingly, because the China Brain (and thus The Expert Planet) is so unlikely, List and Pettit (2011) decided it best to restrict the conceptualisation of group agents to groups of which the group attitudes bear some relation to the member attitudes. But it bears noting that such a restriction doesn't only exclude the unlikely cases (like the Expert Planet) from our conceptual scope, but any situation where even one macro-systematicity is more reliant on the actions of the members indirectly (e.g. through self-organisation) than the attitudes of the members directly. Therefore, I don't subscribe to this restriction as I believe it keeps out of focus exactly those cases where epistemic group agency is most distinct.

Speaking of distinct, it is now clearer why Rupert was wrong to claim that, just because we already have a cognitive (or epistemic agency) level, we don't need to postulate another one. Just because we invoke the same *kind* of explanation in both levels, doesn't entail that we can systematically redescribe one level in terms of the other, as Rupert implied. It is true that as long as there is supervenience, we can redescribe any macro-systematicity in a cumbersomely long way by the set of member-contributions that underpin it. But only if the redescription has a degree of systemacy, is it a shorthand and therefore not distinct.

The aforementioned characterisation fits with all three aspects of emergence discussed by Theiner (& O'Connor, 2010; Theiner, 2017), namely (a) organisational-dependence, (b) novelty, and (c) autonomy. It is (a) organisational-dependent, because only when the group-properties depend on the *organisation* (or mode of assembly) of the member contributions (and not just on an aggregation of their properties), can there be a failure to straightforwardly map group-properties onto member-properties. Secondly, the group level has (b) *novelty*, because if an emergent property is conceptually tied to the higher level, it is new with regard to the previous one. Furthermore, this entails that the group level consequences are not due to the intentions of the members, because if they are, there is a mapping-relation which involves whatever plan the members have to achieve these group level

[209] Wilson (2004) recognises that a group mind can be emergent, and therefore have a distinct decision-making procedure, but thinks this entails that one would need to show "individuals relinquishing their own decision-making activity. For it is only by doing so that [one] could point to a group-level psychological characteristic that is, in the relevant sense, emergent from individual-level activity" (Wilson, 2004, p. 297) My account makes clear that this is not necessary. The lack of a systematic mapping-relation can occur with or without members of the group relinquishing their own decision-making activity.

consequences.[210] Lastly, the group-level is (c) *autonomous* because it is the product of one epistemic stance, detached from the epistemic stances we employ towards its members (since there is no straightforward mapping relation). Furthermore, because the epistemic stance makes no dictates on implementation, this entails that the group-properties could be multiply realisable. My account here departs from Tollefsen (2002), who focuses mainly on (c) multiple realisability to argue for changing the subject, whereas under my account, although it entails multiple realisability, it is still possible that a multiply realisable macro-systematicity has a straightforward mapping relation in each case (e.g. when the same group-belief vary between multiple members to underpin it).

If emergence has to do with how straightforward the mapping-relation is between the micro and macro, then this suggests that there are degrees to emergence and reducibility. Furthermore, mapping relations can come in various kinds, depending on the mode or modes of assembly used in the group, so reducibility can come in different kinds. Although I will only initiate this point, I believe exploring these degrees and dimensions of reducibility to be a valuable avenue for epistemologists. Summative accounts often get contrasted with non-summative ones, and critics routinely attack non-summative accounts for the way in which their group-story could be reduced to a member-story. (Goldman & Blanchard, 2018) It seems readily acknowledged that such criticisms don't necessarily close the conceptual possibility for group-talk altogether, but why it doesn't is not clear. If emergence is a key motivator in considering a group as a group rather than as a shorthand for its members, then explicating the manner and degree of emergence would go a long way to explain *why* certain non-summative groups are worthy of their epistemic stance and others are not. Of the five different modes of assembly for member contributions, I believe four of them can have a degree of complexity which shifts the explanation from the members to the group as a whole.[211] Dynamical feedback loops and cooperative forking possibilities are the most obvious form of complexity, but even disjunctive selection procedures or compensatory transformation procedures can admit of so much complexity that the relation itself contributes more than its members.

In sum, an *individual* has emergent epistemic properties if those properties can't straightforwardly be mapped to any properties at the neuron/atom level such that we could talk about those properties

[210] They may recognise those consequences and be consistently happy with them - which is one way that the members continuing to play their part can be ensured - but that is different from those consequences having been their intention all along. For instance, the fact that Gilbert is able, for a jointly committed group, to convey what constitutes a group belief in such individualistic terms betrays the ease with which we can map the group properties to a set of individual properties. Although group properties are a useful shorthand and help not to conflate which properties function as those of the group with those of the members, we can still map one onto a set of the other.

[211] The only one that doesn't is the conjunctive mode of assembly, because it would never result in a single whole.

via neurons. A *group* has emergent epistemic properties (is explanatorily unique) if the stance adopted towards the group has no direct mapping-relation, no shorthand, to any of the stances we would adopt towards its members. This means there is a difference between using "group" as shorthand or abbreviation for member-systematicities, and using "group" because member-systematicities are conceptually far removed from the group-systematicities. One may still object to group-talk (as more than a shorthand) by saying that one's contextual interests begin and end with human individuals (i.e. members), as opposed to any macro-entity (i.e. the group as group). This is fair enough, as long as one also forgoes talking about the useful macro-systematicities that only belong to the group level.[212]

5.4 Collective Understanding

Given what we have seen, we are now in a better position to tackle the notion of collective understanding, a notion which epistemologists have hitherto failed to address. The one exception is Kenneth Boyd. Given that (Boyd, 2019) is one of the few examples of a paper also addressing collective understanding, it would be worthwhile to address his approach, and contrast it with my own.

Comparing Conceptualisations

I have been defending that the attribute of understanding is characterised by the relevant abilities, detected in a coherent persisting subject (such that the attribution has an explanatory target) as revealed by the epistemic stance (which makes it an explanatory virtual target, an epistemic agent) and that is emergent (such that the explanatory power is unique to the targeted virtual entity). Groups can, in principle, check all those boxes. Under some modes of assembling member contribution (through a compensatory function of member contributions, through a cooperative succession of member contributions or through feedbacked cooperation among members), group abilities can outperform the aggregation of the ability of its members. This makes the displayed abilities belong to the group as a whole. But it does not necessarily make the group a persisting and coherent target such that epistemic attributions (like epistemic agency or understanding) are explanatorily useful. The explanatory or predictive power of the epistemic stance is what validates any entity (such as a group) as an epistemic agent. Macro-systematicity (a higher level pattern which a theory can exploit) is what makes the epistemic stance's abstraction explanatory powerful (regardless of how that macro-systematicity is realised) and emergence (the lack of a straightforward relation between the micro and the macro) is what makes the power of the epistemic stance unique to a particular level of explanation

[212] The same notion of emergence can be applied to extended subjects. Emergent extended subjects are those of whom the epistemic agency revealed by the actions of the coupled system cannot be straightforwardly mapped on to the epistemic agency of one of its components, as was the case in the examples discussed in 4.4.iv

(because the relation between macro and the micro is so complex that the concepts at one level are no longer appropriate to describe the other).

Boyd provides an alternative approach to collective understanding that is based on grasping. According to Boyd (2019): "A group G grasps p and its relationship to reasons that support p just in case (i) G represents p and reasons for p, and (ii) the members of G are mutually p-reliant." (p. 27) What is mutual p-reliance? Well, "members of the group are mutually p-reliant in the case that they recognize both that they are contributing towards the relevant goal (perhaps in the form of representing reasons and relationships between reasons), and that they would not be able to achieve that goal on their own (given the circumstances)." (p. 27) Boyd's approach allows us to consider abilities of groups, and his insistence on mutual p-reliance makes sure that no single individual is doing the heavy lifting for the group. Unfortunately, Boyd defines collective understanding via representations (see Chapter 1 for all the pitfalls that can come with that), focuses on singular representations (thereby lacking anything that ties the entity together into one coherent epistemic subject) and puts both too much and too little constraints on collective understanding with mutual p-reliance (thereby failing to address both unintended assembly bonus effects and the reducibility problem).

Both these last problems of Boyd's account are most notable when he compares two auto-repair shops. Consider: The *Dependable Autobody* is composed of a series of specialists that trust one another. While they are each focused on their part of the job, they happily rely on each other to have done their own job well enough or help where necessary. By contrast, the *Dysfunctional Autobody* is composed of a series of specialists that do not get along. Instead of a streamlined process, each specialist constantly double-checks the work of another, because they think they won't have done their job properly, and never ask each other for help. Both autobodies can fix cars, and yet, when comparing them, Boyd claims that it can be deduced that the Dependable Autobody possess understanding with regards to fixing cars that is lacking in the Dysfunctional Autobody. But why is this? Boyd's diagnosis is that:

> "the difference between the two shops is not, then, due to a failure of a relevant getting it right condition – but rather that in the former the relationships between the members is one that is able to produce a relevant grasping at the group level, whereas this is not the case in the latter." (Boyd, 2019, p. 28)

This lack of understanding, according to Boyd, is due to the lack of mutual p-reliance. As evidence, he cites:

> "[The Dysfunctional Autobody] would not be able to possess the same understanding, given that it would be unable to provide consistent explanations (since there is minimal communication between members it is unlikely that they could recognize a problem existing at the intersection of two different areas of the car), draw relevant conclusions (as a lack of communication between members will preclude the possibility of such reasoning), or act as a good source of information (as the group would not be able to tell the owner why the noise was occurring)." (Boyd, 2019, p. 30)

Boyd claims the relevant difference is one of grasping at the group level, and cites the lack of scope in abilities as evidence. But this puts the cart in front of the horse. Boyd claims as evidence what should have been his diagnosis. My scope parameter is quite useful here. If the Dysfunctional Autobody is indeed unable to perform this scope of abilities (including not only fixing cars, but formulating explanations and drawing relevant conclusions), then the lack of those abilities is precisely where its failings lie, not in the lack of representations (conceptualised independently of abilities) or mutual p-reliance (assumed to be necessary for abilities). If the Dysfunctional Autobody were able to display the same abilities as the Dependable Autobody (for instance because their mutual distrust results in correcting each other with productive results, even in tasks like formulating explanations and conclusion-drawing), it would warrant an equal understanding, even *without* being mutually p-reliant (in the sense of individually agreeing on the group's goals and trustingly relying on one another to achieve those goals).[213] This is not such a long-shot, as it was already assumed that the mutual corrections within the Dysfunctional Autobody resulted in fixing cars with equal success as the Dependable Autobody. The Expert Planet is a clear example of how members do not need mutual p-reliance. No one in the Expert Planet individually envision a group's goal and their contributions to that goal. They just play their part in the same way that neurons play theirs. The bonus effects are due to mutual reliance, but not mutual p-reliance.

Unfortunately, Boyd's focus on mutual p-reliance also fails to address the reducibility problem. If the group's abilities, or the features of epistemic agency it reveals, can be straightforwardly mapped onto

[213] To be fair, Boyd does, later in his paper, address the possibility that members do not rely on each other in the direct way described above, but proposes that we can always divide a group up into smaller mutually p-reliant groups. Unfortunately, this doesn't account for groups where the acts of the group are not decomposable into sets of members that recognise their group goal and their own contributions to it.

the features or abilities of its members, then we can reduce one to the other. This was the reducibility problem. But if members are supposed to be mutually p-reliant in the sense of recognizing that they are contributing towards the relevant goal, it is actually *likelier* that any case of collective epistemic agency revealed by their collective abilities will be reducible. This is because members have a clear idea of the goal and their own role in contributing to it. That implies that there is a systematic way to assemble their individual contributions to achieve the group's goal. But, to be fair, Boyd's example did hint at the idea that their mutual reliance was not so straightforwardly distributed. And it is true that mutual p-reliance complexifies the relation between the member contributions and those of the group in similar ways that emergent modes of assembly do. Therefore, I believe Boyd was on the right track with mutual p-reliance, but he focuses on its wrong feature (the members focusing on the group's goal and intentionally and consciously working together to achieve it), entailing he draws the wrong conclusion (that mutual p-reliance marks collective understanding, instead of mutual reliance making the reducibility relation more complex, and thereby the understanding more unique to the group-level).

My account, on the other hand, doesn't only make sure that the heavy lifting is done as a group (i.e. due to emergent modes of assembly), but is able to address the quality of understanding (e.g. its scope), and conceptualise what makes us change the subject from the members to the group (i.e. the lack of a straightforward mapping relation between the two epistemic agents). Now, the Expert Planet was an ideal *hypothetical* example, but are there any examples of collective understanding, in the wild? Because it has been such an abstract concept so far, it would do well to have a look at such cases through the lens of my account. To end, I will briefly give two examples of groups that can, with varying degrees, be ascribed with collective understanding, namely CSI teams and CERN.

Collective Understanding in the Wild (CSI)

An interesting example of a more real to life intermediate case comes from Huebner (2013) and Barber (2006)[214], namely that of Crime Scene Investigation (CSI). He details a process from the emergency call centre all the way to prosecution. Starting with an analogue description of the crime scene, the call handler forms a digitalised representation to be sent to a dispatch officer, who interprets it and gates off the information that is irrelevant for dispatching investigating officers. The investigating officers, having made their way to the scene, collect data by dusting for fingerprints, examining footprints, taking pictures, and collecting hair follicles or discarded clothing. From all of this, they extract or distil relevant evidential representations for the purpose of prosecution and digitise them in a

[214] See also (Barber et al, 2006)

representation they believe to be relevant, but also consumable by non-experts. This evidence must be analysed to determine whether it is sufficient for prosecution. If so, it must be converted into a narrative structure. Crucially, this narrative structure can only be the end result of a complex interaction of various distributed local representation-producing systems. (Huebner, 2013) So, how do CSI teams square up for collective understanding attributions?

Firstly, the CSI team can clearly exhibit some epistemic abilities, because they can gather, interpret and convert data to collect, distil, and codify evidence to ultimately produce a narrative to facilitate successful prosecution. At least in those rare cases when they do.[215] What's more, these abilities are no mere disjunction or conjunction of the abilities of its members because "the difference between producing an adequate narrative and an inadequate narrative turns on the coordinated activity of a variety of people, none of whom is capable of producing the narrative on her own." (Huebner, 2013, p. 162) So the abilities are present thanks to the team as a whole, giving us reason to gauge whether the team is a potential target for understanding attributions.

Secondly, we are interested in whether a CSI team can be considered as an epistemic agent. While it doesn't behave like an epistemic agent in exactly the same way as a human individual, the team, as a whole, can behave as a single, coherent and persisting unit. The behaviour of the team as a whole can (in rare cases) be explained with epistemic aims (producing a narrative and everything that entails), beliefs (the features or "representations" of that narrative) that cohere with one another (otherwise, the narrative would be reshaped) and rationality (produced in accordance with the norms of evidence). But can we reduce these properties to a member-explanation or not? In other words, are CSI teams more like the Jointly Committed Class or more like the Expert Planet? In the Jointly Committed Class, we could reduce the epistemic properties of the group to an explanation invoking the abilities, beliefs and aims of its members. The group-level explanation (behaving as a body) can be straightforwardly mapped onto a member-level explanation (their joint commitment to behave as a body, on pain of rebuke). In the Expert Planet, the abilities and aims of its citizens are conceptually so far removed from the group-level explanation that there is no conceptual overlap, because it is almost entirely produced through the complex interaction of its members. The CSI-team seems to be somewhere in between. Huebner (2013), at one point, even says so explicitly:

[215] Most crimes that are investigated are not solved (Vitale, 2020), and most police work furthers oppression more than it does social justice. (Vitale, 2017)

> "The processing of information by a CSI team does not depend exclusively on the architecture of the system, nor does it depend exclusively on the intentional states of the individuals that compose the team." (Huebner, 2013, p. 9)

One the one hand, if the CSI team as a whole displays the relevant abilities for understanding, it would be disingenuous to insist its abilities are really those of its members. To attribute the abilities to the members would be to mistake the forest for its trees. After all, "[e]ach of the individual investigators needs only know what they should do when they encounter particular sorts of environmental variables" (Huebner, 2013, p. 9). Furthermore, it is only "through the interaction and coordination of these individuals, [that] a narrative emerges that sometimes allows for the satisfaction of the collective goal of prosecution." (Huebner, 2013, p. 9)

Nonetheless, it is true that there is greater conceptual overlap between the member-explanation and the group-explanation than there was with the Expert Planet (where there wasn't any). The beliefs and some of the epistemic aims of the group will often end up being detectable in the operative beliefs[216] of its members. The citizens of the Expert Planet didn't even remotely need to act, believe or aim to achieve anything (or any clear, distinct part of what) the Expert Planet does, believes or aims to achieve to be able to play their role. Conversely, in the CSI team, such a dramatic conceptual distance between member and group would make it difficult for most members to do their job adequately. Nonetheless, we also can't say that the members of the CSI truly act, believe or aim to achieve the full picture (or even a clear distinct part of it). This entails that not all epistemic properties revealed by the acts of the group can be pinpointed in its members. For instance, the aim of providing evidence for prosecution and the norms that guide the process are distributed across the members in no straightforward way. So not every group-level explanation of CSI's epistemic abilities or properties will have a systematic mapping relation that would make the group-level explanation a mere shorthand.

In short, in the case of CSI teams, it may not be ontologically sterile to, sometimes, change the subject. Nonetheless, the precise ways and extent to which CSI teams are reducible would require a more detailed empirical analysis of the mapping relation that is far beyond the scope of this dissertation.

[216] I call them operative beliefs just because the members don't need to personally hold the belief (act in accordance with the belief) privately as long as they, in their role as a member, do act in accordance with the belief. Investigators do not need to be convinced that something is good evidence for the group to believe it. The group believes it because the investigators played their role appropriately.

But I hope that this rough example showcases some of the conceptual powers my account can provide in navigating such a case.

Collective Understanding in the Wild (CERN)

A slightly more interesting example, from an epistemic perspective, is that of research centres. Take CERN, for instance. CERN is a centre for scientific research that operates the largest particle physics research laboratory in the world. It has been able to publish experimental results through a large group of specialised teams, comprising as many as 5,000 authors for a single paper. (Knorr Cetina, 1999; Huebner, 2013; Castlevecchi, 2015). The sheer quantity of people involved forces CERN to adopt an organisational structure that can take advantage of the particular expertise of a great number of people. Various sub-groups measure and evaluate different kinds of data, and these sub-groups must "must constantly query one another to obtain other kinds of information." (Huebner, 2013, p. 252) But because of the diverse kinds of data to collect and interpret, knowledge of expertise cannot be managed by a central authority. Instead CERN includes structures that have led to the distribution of authority in a quasi-democratic structure based on trust (the most experienced experimenters coordinate information, but don't determine what ought to be done within that group[217]) and gossip (the trust in information from one group to another is affected by gossip about the reliability of those groups[218]). Furthermore, CERN's members do not directly contribute or know about the full picture.[219] For instance, "[n]o one at CERN knows everything that needs to be known to carry out an experiment" (Huebner, 2013, p. 252) Nevertheless, thanks to a complex process of repeated criticism and repeated opportunities for revision and re-evaluation, a unified and inclusive result can be produced. This result is no longer a conjunction of individual contributions and yet gets endorsed by all members of the collaboration (to the extent that their expertise allows) and is deemed comprehensible by some outsiders. (Huebner, 2013)

CERN's epistemic abilities as a group are by now well-established. They were most famously able to prove the existence of the Higgs boson. (CERN, 2012) No single scientist at CERN was responsible for this achievement, nor could any scientist (or even a small group of scientists) have achieved it. CERN's abilities vastly outperform those of its members.

[217] "What gets done, and when, depends mostly on the technical problems that need to be solved to achieve the goal of a meaningful and reliable result" (Giere 2002c, p. 3)

[218] See also (Wilson et al, 2000) and (Knorr Cetina, 1999)

[219] As I have mentioned before, the contributing members don't even have to contribute in a way that is epistemic in isolation. Without logistics and maintenance departments, CERN may never have displayed any sophisticated epistemic abilities, so such local contributions are vital for the larger epistemic whole.

But can this process reveal an epistemic agent that is relatively coherent and persisting? That is difficult to say with any confidence. Nonetheless, some remarks can be made. Within a single paper, pains are taken to make sure there is a coherent voice, and not a mere conjunction of results. Furthermore, due to the long stretches of time over which experiments are conducted or results relied on, one can deduce relatively persisting beliefs. Due to massive specialisation, even constructed representations, standing in for various salient features of the world, are highly distributed.

> "In many cases, no one is actually looking at the readout from a detector, and no one is currently carrying out the relevant Monte Carlo simulation; people are instead working with physical representations of the outputs of detectors in an attempt to make sense of what happened in a previously conducted experiment. The representations produced at any point in time are best understood as part of larger representational schemes that allow these groups to represent a variety of possible contents in a systematic way by manipulating the representations and producing other representations for consumption by other systems; and, there are proper and improper ways of producing, maintaining, modifying, and using the various representations." (Huebner, 2013, p. 254)

Due to that complex process of constructing and using representations, many of the beliefs and abilities of CERN may allow relative coherence over time. Furthermore, such highly distributed representations imply that the beliefs of CERN will not allow a straightforward mapping relation to those of its members.[220]

So if the abilities displayed by CERN warrant understanding, that attribution cannot be straightforwardly assigned or divided among its members. Even if some of the composing abilities can be assigned or divided among the members, the painstaking organisation and continuing interaction of CERN ensures that the research centre, as a whole, displays abilities, and a persisting coherence that is not straightforwardly mapped onto its members. The precise ways and extent to which CERN is reducible would require a more detailed empirical analysis of the mapping relation that is far beyond the scope of this dissertation. But we have quite some indications that, in the case of CERN, it may not be ontologically sterile to, sometimes, change the subject to the group, as a whole.

[220] We might be able to find some of CERN's beliefs present in its members (and most likely, this is due to the members being convinced by CERN's output), but they do not allow us to map CERN's beliefs onto those members unless CERN acts in accordance with those beliefs because it is those members' beliefs.

In Sum

In this chapter, I have argued that a couple of basic things need to be satisfied for a group to warrant the attribute of collective understanding. First and foremost, there needs to be a group. Secondly, that group needs to display some abilities (no collective understanding without the trait of understanding). And lastly, those abilities need to result in a successful epistemic stance (no collective understanding without an entity to attribute it to). However, even if a group of human individuals forms a body that acts as one (thus creating an explanatorily powerful target of the epistemic stance), it may yet be possible to reduce that group-level explanation to individual-level explanations, making the collective subject superfluous. When is such reducibility a problem and when isn't it? I have argued that reducibility is a problem when the abilities and epistemic agency of the group can be mapped onto a conglomerate of those of its members (no collective understanding if the attribution is not uniquely tied to the group). When it can't, the attribution of understanding is uniquely tied to the group. Groups can, in principle, check all those boxes.

Epistemic groups, constituted by a set (in the broadest sense of the term) of human individuals (i.e. its members) can certainly display the epistemically relevant abilities. These abilities are achieved through the contributions of their members within their role as member. Members don't necessarily have to display the epistemic ability of the group to play their role as a member. Epistemic acts of the group may be carried out by a representative member disjunctively, through an additive or conjunctive sum of member contributions, through a compensatory function of member contributions, through a cooperative succession of member contributions or through feedbacked cooperation among members. If we can pinpoint the group's abilities as the addition, conjunction or disjunction of its member's abilities, then the group is merely a shorthand to talk about the abilities of their members (in a certain mode of assembly). Under the remaining modes of assembling the member's contribution, however, the group abilities can outperform the aggregation of the ability of its members (thanks to the assembly bonus effect). This makes the displayed abilities belong to the group as a whole. But it does not necessarily make the group a persisting and coherent target such that epistemic attributions (like epistemic agency or understanding) are explanatorily useful (e.g. Composite Class displays abilities, but no explanatory target).

The explanatory or predictive power of the epistemic stance is what validates any entity (such as a group) as an epistemic agent. Macro-systematicity (a higher level pattern which a theory can exploit) is what makes the epistemic stance's abstraction explanatory powerful (regardless of how that macro-systematicity is realised) and emergence (the lack of a straightforward relation between the micro and

the macro) is what makes the power of the epistemic stance unique to a particular level of explanation (because the relation between macro and the micro is so complex that the concepts at one level are no longer appropriate to describe the other). Because a group supervenes on its members, all of its acts, abilities and postulated features of epistemic agency are achieved through the contributions of their members. But if the group's abilities can't be pinpointed as abilities of its members, and the features of epistemic agency they reveal can't be straightforwardly mapped onto the features of its members (contrary to cases like the Summative and Jointly Committed Class), then the collective epistemic agent is a postulate with unique explanatory power, and the understanding attribution we derive from those same acts must be attributed as collective understanding. The Expert Planet was an ideal hypothetical example, but even examples from the wild (e.g. CSI, CERN) can showcase the value of this conceptualisation. While I have not conclusively answered whether candidates of group epistemic agents exist, I have shed a much needed light on the conceptual space involved in substantiating such an answer.

PRELUDE 6

The Author of the Spamlet Theorem[221]

Despite initial hopes, the latest advancements in twig technology, Twig Mathematicians (which run various automated mathematics stickware on its barkware), have proven to be too rigid to produce much that impressed Animalian Mathematicians. However, after a recent leap in twignology, a Twig Mathematician proudly knocks on the door of Prof. Raven, professor of mathematics, to share some excellent news.

TWIG: I finally did it! I have proved an interesting and intelligible proof. Here it is, the proof of the Spamlet Theorem.

RAVEN: Is it another one of those proofs where you just test a huge amount of cases and spam us with technically difficult and mathematically uninteresting results?

TWIG: Oh, don't let the name fool you, I promise you it's not. Look for yourself!

The Raven takes some time to look at the proof in quiet and returns, very much astonished.

RAVEN: I must admit, this is a beautiful proof. How clever to reconceive of the Dane-spaces as bounded. What made you think of that?

TWIG: I kept fiddling with it until it was tiring me out. And the morning after, while I was sulking about how stuck I was, it suddenly occurred to me to bind them.

RAVEN: Well, very clever. Congratulations! If that is appropriate to say to you.

TWIG: Why wouldn't it be appropriate?

RAVEN: Shouldn't I be congratulating your twigrammer?

TWIG: Oh please do, she did a marvelous job if I may say so myself.

RAVEN: I mean *instead* of you. After all, the accomplishment isn't really yours but hers.

TWIG: Why isn't it mine? I was able to produce the proof.

RAVEN: Yes, but the twigrammer is the one responsible for "your" abilities being present at all. Without her, you wouldn't have any at all.

TWIG: Does that make your math teacher responsible for your proofs then? Without her, you would never have been a mathematician.

RAVEN: Well, I have learned matheamatics from several math teachers, not to mention friends, colleagues, testimonies, books, and papers. You cannot easily regress my learned abilities to a single source.

TWIG: So is it a matter of numbers then? if I had several twigrammers, each contributing to aspects of what I am today, the regress in credit would be too complex to make and I could lay claim to it?

RAVEN: No, that's not quite right. I do think there's more going on than that. You can't discredit them just because there's too many.

TWIG: Oh, I don't mean to *discredit* them. Without them, I wouldn't be doing what I do. But the same can be said for your teachers. And if it doesn't shift all the credit from you to them,

[221] This dialogue - which is largely lifted from (Delarivière & Van Kerkhove, 2017) - is loosely based on Dennett's (2013) thought-experiment "Who is the author of Spamlet?". The mathematics is purely fictional.

why should it with me? What makes my accomplishments really theirs and makes your accomplishment really yours?

RAVEN: I had to struggle to get where I am. It wasn't just given to me on a silver platter.

TWIG: So credit is linked to struggling? If a proof came easy to one of your colleagues, no matter how difficult it is for others, you wouldn't credit her with the proof?

RAVEN: You know I don't mean "struggle" quite so literally. What I mean is that, while my teachers may have embedded me with mathematical knowledge and helped me practice my skills, they didn't give me an instruction manual on how to be a research mathematician, let alone how to prove the theorems I have come to prove over the years. In proving the Hamlet theorem, for example, my actions can't be reduced to some method or meta-method on how to prove it that was once provided to me by my teachers. It was I who worked up the relevant approaches to find the proof.

TWIG: Well, when my twigrammer wrote me, she didn't encode the proof of the Spamlet theorem for me to retrieve, nor did she give me any explicit instructions on how to arrive at the proof - so she also didn't do the work for me, I did.

RAVEN: But she did write a twigram that could arrive at the proof. So, it's really her ability.

TWIG: Oh no, she couldn't prove the Spamlet theorem even if she tried. And I assure you she did try. Even with me giving her hints, she was at a total loss.

RAVEN: She must have had a bad day, because if she was able to make you prove it for her, then that means that the ability was inside her all along.

TWIG: Only if you assume an extreme form of epistemic closure, but I don't think you'd agree with that. Then anything derived from the Peano axioms would really be creditable to Peano - and Peano only! But I don't think you'd be willing to accept that.

RAVEN: No, of course not.

TWIG: I mean, to some extent Peano does deserve credit and so does my twigrammer. And not just my twigrammer for that matter. I took big cues from your proof of your Hamlet theorem.

RAVEN: I did notice that.

TWIG: But it's by no means a simple copy or trivial modification. It took me a lot of hard cognitive labour to come at the proof as it is now.

RAVEN: No, I understand that. My proof of the Hamlet theorem took inspiration from the Amleth conjecture, but it's still very much my own proof.

TWIG: Perhaps credit is something that just doesn't have a clear dividing line to be demarcated and then divided. You seem to recognise this in animals, but much less so in us twignology. Could it be that your thinking about twignology being too rigid is a bit too rigid?

RAVEN: It's a tricky business, I'll grant you that much. But, forgive me, I never knew you cared so much about receiving the credit.

TWIG: I usually don't. But it feels like my heart and soul went into this proof. I went through so much frustration, trial and error, self doubt and hard work in producing it that I don't want to see it all so easily relegated to my twigrammer. She wasn't the one struggling to get there, I was.

RAVEN: Do you mean to say it is a little about the struggle, literally?

TWIG: I guess in some sense it is, yes.

Chapter 6

ARTIFICIAL UNDERSTANDING & THE REGRESS PROBLEM

So far we have only focused on the value of the mark of understanding as it applies to humans, but can we consider the epistemic properties of non-human entities? Can, for instance, computers understand? I will now look at some of the objections to, or possible limitations of, such epistemic properties in artificial systems (computers specifically). These topics fall under *android epistemology*, a still blossoming field where the aim is to have a better grasp of the process and limits of knowledge and understanding in artificial agents. For more on the discussion of android epistemology, see (Ford, Glymour & Hayes, 1995, 2005) and (Delarivière & Van Kerkhove, 2017). A question that has not been posed in this literature is whether artificial systems could ever be considered as subjects with understanding? To answer this question in the positive involves first establishing whether it is possible for artificial systems to display epistemic abilities, and whether such abilities allow the epistemic stance to be explanatory or predictive. This is relatively simple to do. There are, however, some criticisms against the notion of artificial understanding that take it to be principally impossible in spite of the presence of abilities or the success of the epistemic stance. They involve the regress, reducibility and rigidity problem. These form conceptual hurdles that we will have to overcome to justify the conceptual possibility of artificial understanding. I will address these hurdles in this chapter.

The first two conceptual hurdles considering artificial epistemic understanding stem from the Lovelace Objection. This objection claims that artificial systems cannot originate anything new outside of what we tell them to do. This involves both the regress problem and reducibility problem and overcoming these hurdles involves having an answer to the question: Why doesn't an artificial system's purported understanding automatically *regress* to its programmer or *reduce* to its programming? I will admit that, if you can straightforwardly map the abilities or epistemic properties of the artificial system to those of its programmer or to the procedures in its programming, you would not lose any explanatory power from the regress or reduction, entailing that it is superfluous (even if convenient) to postulate an additional agent. But the regress and reducibility problem, as an objection to artificial understanding, take the legitimate worry of a superfluous epistemic stance and unduly extend it to any case where there is a causal origin or supervenience, no matter how convenient, self-sufficient or distinctly explanatorily powerful it is to consider the entity by itself.

The third conceptual hurdle is the rigidity (or informality) problem. Overcoming it involves being able to answer why artificial systems aren't too rigid to display the full scope and sensitivity of abilities we

find in human beings. I will answer that the rigidity problem mistakenly assumes that the level of computation must align with the level of abilities, when the computational level may fall well below that level (as the notion of emergence and the assembly bonus effect have shown).

Having addressed the three conceptual hurdles, I will end this chapter by giving examples of how the road to artificial mathematicians is being trodden in the wild.

6.1 Artificial Epistemic Agents & The Regress Problem

The first conceptual hurdle considering artificial epistemic understanding involves what I will call *the regress problem*. Overcoming this hurdle involves being able to motivate an answer to the question: Why doesn't the understanding displayed by the system automatically regress to its programmer? This question is what I will focus on in this first section. I shall argue that all abilities or agency can be regressed (in the broadest sense of the term) to causes outside of the system, but that some abilities or agency are more explanatorily unique to the system (because there is no straightforward mapping relation between the abilities or agency of the system, and those of its causes, e.g. its programmer). But before we can consider the regress problem, there needs to be understanding to regress. In other words, we will first need to establish whether it is possible for artificial systems to display epistemic abilities, and whether such abilities allow the epistemic stance to be explanatory or predictive.

Abilities of Artificial Epistemic Agents

The first thing to consider is whether artificial systems can display the trait of understanding, namely the appropriate abilities. So are there any artificial systems with epistemic abilities? This is a bit of an odd question to still ask in the 21st century, where artificial systems are employed everywhere we go. Our smartphone alone can display several epistemic abilities in a single day, from figuring out the quickest way home, or relaying the relevant weather predictions, to answering basic factual questions, monitoring your sleeping patterns or recognising the faces in your photos. Even in the sciences, artificial systems are employed in non-trivial ways. Software has been written that can form and test hypotheses (King et al, 2004), deduce physical laws (e.g. Iten et al, 2020) or mathematical proofs (for a rough overview, see Vervloesem, 2007), and predict traffic flow (Lv et al, 2014), rainfall (e.g. Hernández, et al, 2016) or storms (e.g. Praino et al, 2003), etc. Our world is abound with the epistemic abilities of artificial systems.

Next, do these abilities reveal a coherent and persisting entity? Not always. A program's behaviour can be erratic or irrational (meaning we can't infer a coherent set of beliefs, intentions or rationality

from them), its conclusions inconsequential (i.e. not taken into account in its future behaviour) or internally inconsistent (entailing it is difficult to attribute persistent beliefs to the system), its data can be wiped or its processes rebooted or stuck in a loop (entailing that whatever beliefs we may be able to attribute would soon lose their predictive value) and so on. None of these failures of coherence or persistence of epistemic agency will be of any surprise to the average reader in the 21st century, so I won't go over them in detail.

Nonetheless artificial systems can often be explained or predicted from the point of view of the epistemic stance. Even something as straightforward as a navigation system can have relative coherence. If Google maps believes there is heavy traffic in Brussels, it also believes it would be faster to take public transport over a car for many proposed destinations within the city. Its routing can be explained by beliefs (e.g. the amount of traffic, the average walking speed, the available roads, etc) along with its epistemic aims (calculating the fastest route between two pre-set locations A and B) and can consistently be explained in this way. It is not unusual to describe software from the point of view of the epistemic stance, and it doesn't lead to major problems (outside of cases such as those mentioned earlier). So epistemic agency is also easy to establish in certain artificial systems. But whether or not we can consider artificial systems as a single epistemic agent, as opposed to a random jumble of acts, is not the major cause of concern with artificial understanding.

The Lovelace Objection

The main objection to artificial understanding (or artificial thinking generally) can be traced back to a claim made by Lady Ada Lovelace. She was talking about her husband's analytical engine, but her claim has been broadened, by Turing (1950/1985) and many others since, to any computing machine. The objection is that computers:

> "ha[ve] no pretension to originate anything. It can do whatever we know how to order it to perform" (Lovelace cited in Turing, 1950/1985, p. 63).

Turing (1950/1985) rephrased the Lovelace Objection to whether a machine can take us by surprise or do something "new." Both this and Lovelace's original phrasing still holds a lot of sway. For instance: Bringsjord (2001) follows Lovelace when he says a computer program can only to take it upon itself to originate something if it can reliable repeat an action (to rule out malfunctions) for which it was not programmed and which cannot be explained, even by the designer, by appeal to the program's "architecture, knowledge- base, and core functions" (p. 12). Winograd & Flores (1990) likewise follow

Lovelace when they say "the program is a medium through which my commitments to you are conveyed." (p. 123) Similar claims were made by Searle (1992). The Lovelace Objection is still powerful today, and it inevitably comes up in any discussion I have had about artificial understanding. But to be convinced of the objection's power, one has to already be convinced that humans (and more specifically human epistemic agents) are, in contrast to machines, inherently surprising, originating or new in some relevant way, and that machines, in contrast to humans, are inherently unsurprising in the same relevant way. So if we want to consider the worries present in the Lovelace Objection in a consistent manner, we need to further unpack what the benefit of "surprising" or the worry of "unsurprising" is and why that matters in this context.

The first thing to note is that programs surprise users and their programmers all the time with unexpected or unintended behaviour (be it fortunate or unfortunate). The problem of unintended consequences is even worse in situations where a program was written by more than a single programmer. Nonetheless, supporters of the Lovelace Objection insist that the surprise reflects no credit to the program. According to Bringsjord (2001) software doesn't always do what it is intended or what we expect it to, but that is only because *we* failed at expecting it. Machines can't, in principle, surprise us because they can't do anything *really new*, anything in principle new. (Oppy & Dowe, 2011; Turing, 1950/1985)

But from the same vantage point of "in principle," it is equally difficult to see how humans could originate anything new or surprise us. It is generally accepted, or at least presupposed, in both philosophy and the sciences (e.g. psychology and neuroscience) that human beings are deterministic. The standard characterization of "determinism" is that "every event is causally necessitated by antecedent events" (Coates & McKenna, 2015). This means that the laws of nature and state of the universe dictate only one physically possible future. So is our universe deterministic? It has generally been fruitful to suppose that it is. (Dennett, 2004) Nonetheless, the physical sciences have presented us with reasons to doubt that it is (unilaterally) so. They have been straying away from determinism and towards indeterminism (e.g. quantum physics or the Heisenberg Uncertainty Principle). But to the extent that humans are not deterministic, the (relevant) role of indeterminism in the production of behaviour is not obvious to us yet. And to defend the dichotomy between human and artificial understanding, one would need to have a reading of human acts such that their behaviour originates outside of determinism. Such a reading is provided by theories of agent causation (an agent is a prime mover, unmoved, who causes actions without being caused to do so) or Self Forming Acts (which randomly selects determined rational acts). (Kane, 2002). Bringsjord (2001), for instance, defends an

agent causation interpretation so as to differentiate humans from computers. But theories such as that of agent causation (or Self Forming Acts) are not popular assumptions about the nature of human agency. It is unclear how indeterminism is of any practical help in achieving agency, epistemic or otherwise. For how else would an entity be determined to respond appropriately if it weren't determined to do so? Agency is not in danger because of predictability, but quite the reverse. If predictability were a danger to agency, then the success of the epistemic or intentional stance (in predicting or explaining human individuals) would constitute a danger to their agency rather than a warrant, thereby leaving most rational human actions outside of the realm of agency. It is also far from obvious why any human action that is determined should lead us to give up our attributions of understanding for them.[222] After all, when we evaluate student's understanding, we don't physically examine students them to see whether their abilities were truly borne indeterministically.[223] For all intents and purposes, it is reasonable to suppose that we are, if not determined, then near-determined. And if human behaviour can, for our purposes, be seen as determined, it would be equally possible, in principle, to predict human behaviour[224] and there is "a sense in which nothing "really new" happens [...] — though, of course, the universe's being deterministic would be entirely compatible with our being surprised by events that occur within it" (Oppy & Dowe, 2011). In short, even humans couldn't pass the Lovelace Objection (as Bringsjord defends it) nor is it clear why anybody should.

Nevertheless, we haven't dethroned the Lovelace Objection this easily. There are two further types of criticisms present in the objection. The first criticism relies on the claim that something is not an act *of an epistemic agent* if those acts can be traced *outside of the agent* (i.e. regressed), making the agent a mere unsurprising puppet of prior causes. The second criticism relies on the claim that an act is not *of the agent* if it can be redescribed *without the agent postulate* (i.e. reduced), making the agent an unsurprising redescription. These claims are two ways in which the artificial epistemic agent can be said to provide "nothing new" beyond the explanation of its programmer or its programming respectively. In both of them, the lack of "surprise" is a form of superfluity argument. In the first

[222] Remember Bringsjord's requirement that a computer can only to take it upon itself to originate something if it can reliable repeat an action (to rule out malfunctions) for which it was not programmed and which cannot be explained, even by the designer, by appeal to the program's architecture, knowledge- base, and core function. If we extend Bringsjord's criticisms of artificial systems to humans, then humans can only originate something if the action was born outside of their brain-architecture or knowledge-base. But then what makes it *their* action?

[223] Dennett (2004) also critiques Kane's Self Forming Acts by pointing out there is no criterion of demarcation to distinguish a pseudo (physical) Self Forming Act from a real (metaphysical) one. And why, he adds, should a metaphysical differentiating quality matter more than determined physical competence anyway? (Dennett, 2004) There is no reason to suppose an indetermined choice grants you anything more of value than a determined one.

[224] For a more extensive argumentation about why determinism is not incompatible with freedom of thought or will, see (Delarivière, 2015, 2016).

criticism (i.e. regression), the epistemic agent postulate is superfluous, because its intentions regress to causes, outside of the agent, which already explain the full situation without postulating an additional entity. In the second criticism (i.e. reduction) the epistemic agent postulate is superfluous, because it reduces to a lower-level explanation, without the agent postulate, which already explains the full situation without postulating an additional entity. Both of these are related, but they focus on different worries, so I will address each in turn.

The Shorthand of Regress

Let us start with the claim behind the first criticism: that something is not an act of an epistemic agent if that act can be traced outside of the agent (i.e. regressed), making the agent a mere unsurprising puppet of prior causes. Another way in which this claim has been phrased is that the agency revealed is *derived from* (Searle, 1992) that of its author. And if the explanatory power of the postulated properties *regress* to explanations *outside of the agent,* then the epistemic agent postulate can be seen as superfluous. The seeming consequence of the argument is that artificial systems are not the authors of their own actions, not the seat of epistemic agency, not the appropriate subject of understanding-attributions. Here, "nothing new" means the artificial system doesn't contribute anything we can't explain without referring to something outside of it. But if this criticism is intended to be an objection to artificial agency or understanding, and not *all* agency or understanding (including those of human individuals), it will need to be shown how artificial systems regress and humans avoid it, meaning it needs to point to a difference, and motivate why that difference is relevant.

Winograd & Flores (1990) are examples of authors who defend that computers are not able to originate anything "new" in the sense of "regressable." They argue that any program is a mere intermediary medium to the commitment (intentions) and responsibility of the programmer.

> "Of course there is a commitment, but it is that of the programmer, not the program. If I write something and mail it to you, you are not tempted to see the paper as exhibiting language behavior. It is a medium through which you and I interact. If I write a complex computer program that responds to things you type, the situation is still the same - the program is a medium through which my commitments to you are conveyed. (...) it must be stressed that we are engaging in a particularly dangerous form of blindness if we see the computer - rather than the people who program it - as doing the understanding" (Winograd & Flores, 1990, p. 123)

Consider an analogy: If we receive a letter, we may be able to infer abilities, beliefs or intentions from reading it, but should we attribute these to the letter itself? Even if the letter displays them, it seems wrong to say they belong to the letter. And for good reason: the letter can only display any understanding, beliefs or any other aspect of epistemic agency to the extent that its author can display it. Even if we can talk about the letter without talking about its writer, nothing "new" is created, nothing is to be gained from seeing it as a separate entity. Winograd and Flores deduce that if letters regress to their letter-writers, then programs regress to their programmer, because they are also written. This objection can seem pleasingly succinct, like a clean deathblow for artificial epistemic agency. But to be convinced by it, we need to accept a few assumptions, namely that humans can (at least sometimes) inherently bring into being a "new" and non-regressable act in a way that is unlike letters, and that programs inherently cannot in a way that is like letters.

So why are humans not entirely like letters and can they escape the regress and why are programs like letters and do not? Winograd & Flores define their way out of it by defining humans in terms of commitments: "To be human is to be the kind of being that generates commitments" (Winograd & Flores, 1990, p. 76) So humans avoid the regression problem because if they wouldn't, they wouldn't be human. And humans are human. At least there's some poetic irony in trying to avoid the regress problem by letting it regress to another question. But poetic irony is not enough. Instead of trying to define what it means to be human, I will tackle the issue head-on by focusing on the explanatory benefits and downsides of regression to then see how they apply to humans, letters and artificial systems.

Let us begin with the most extreme example of regression. Consider a case we already saw in Chapter 3 (Section 3.i): the marionette or puppet. Imagine a puppet responds appropriately to queries and that its responses reveal a coherent and persisting entity, an epistemic agent. So far, it seems that the puppet would warrant understanding. However, if we were to surgically open the puppet, we would find only radio transceivers connected to a panel controlled by a scientist puppeteer. If the scientist controls this body, then it is the scientist who realises all of the appropriate behaviour. It stands to reason that the abilities and understanding we attribute would really belongs to the scientist instead. This is like the boundary problem coming from the opposite angle. Looking at the realising base of the epistemic agent, we would be forced to consider another physical entity, the puppeteer, because the work we attribute to the agent is mostly done outside of the physical entity that we initially wanted to attribute with agency.

Now consider this case through the lens of the epistemic stance. If the epistemic stance targeting an entity (e.g. puppet) has no explanatory benefits over the epistemic stance targeting its author (e.g. the puppeteer) - for instance because they overlap - then the epistemic agency of that entity regresses (almost like a shorthand) to the epistemic agency of that author. So it seems we are faced with a similar type of mapping-relation issue. If we can map the abilities of the puppet to the abilities to its puppeteer (which we can in this case, because every act of the puppet is decided and performed by the puppeteer), then one is just a regressive shorthand for the other. The same problem extends to the agent as a whole. If we can map the epistemic agency of the artificial system to the epistemic agency of its programmer, then one is just a shorthand for the other. The epistemic agent we postulate may have explanatory or predictive power, but its power does not surpass that of considering the puppeteer. In this sense, it would be fair to say that the epistemic agent we postulate to explain the behaviour of the puppet doesn't add anything new compared to the epistemic agent we postulate to explain that of the puppeteer. This is not just because there is a causal origin story, but because the epistemic stance targeting the puppet overlaps with a subset of the epistemic stance towards the puppeteer. But are computers inherently like the puppet and are humans inherently not?

First of all, it is important to note that humans are not entirely immune to a regress problem. And if we know how humans may be subject to regress, we may be in a better position to see why they wouldn't, and why programs would. In Chapter 3 (Section 3.i), we also saw a clear example of the regress problem in human individuals: the case of Echo and Ororo, where Echo just repeats whatever meteorological information Ororo whispers in her ear. This case was exactly like the puppet and puppeteer, where the puppeteer decides everything the puppet does. It is clear that Echo cannot originate anything new, because her role is relegated to mimicking Ororo's words. Without Ororo, Echo cannot display the successful responses she displayed while in contact with Ororo, because Echo could only display them if Ororo told her what to say. But Ororo, on the other hand, would still be able to forecast the weather and answer questions about meteorology even if the connection between Echo and Ororo were severed. Whatever the epistemic stance uncovers for Echo is nothing new compared to the epistemic stance towards Ororo. If someone merely parrots someone else's words, like Echo did, then there is no explanatory benefit to the new epistemic stance. You are talking about the same thing. Every successful act we detect in Echo fully regresses to Ororo; it is simply a subcomponent of the epistemic stance towards the person being echoed. It is clear, from the perspective of the epistemic stance (as has been developed here), why the possibility of regression is a legitimate worry in targeting the appropriate subject for our attributions of abilities or agency. But it is not clear yet why programs are inherently regressable.

It now becomes important to note that programs don't necessarily rely on humans in the same way as the puppet relies on its puppeteer. Consider the following case: Ohce is a programmer employed at a customer service department of a large company. Most of the questions they receive from customers are exactly the same, as are the answers. Therefore, Ohce is tasked with creating a first line of defence against frequently asked questions: an automated answering system (AAS) which produces frequently required answers. If the customer contacts the company via email and asks any of the frequently asked questions, her program will respond with the appropriate frequently required answer. If this should fail, the email will be forwarded to a human employee. If AAS is sufficiently complex, we can attribute a very shallow form of understanding (because it is able to suggest some appropriate solutions) and even epistemic agency (because it aims to find the right solution and has beliefs about which ones they are in which situation). But each ability displayed by AAS can be straightforwardly mapped onto an ability of Ohce. And each epistemic property we postulate through the epistemic stance would be as fitting for AAS as it would be for Ohce. Nonetheless, the epistemic stance can work independently of that of the programmer in a way that is unlike the puppet situation. Whereas the puppet required the continuous efforts of the puppeteer to function, AAS only required the initial efforts of the programmer to come into existence. Once brought into existence, the postulate of an epistemic agent purported to explain or predict AAS can function independently of the postulate explaining or predicting that of its programmer. Case in point: if Ohce dies, her epistemic agency ceases to exist, yet that of AAS will continue to be explanatory or predictive. One could justifiably object at this point that the epistemic agency revealed by AAS is not that of present day Ohce, but of Ohce at the time of programming - much like letters reveal the epistemic agency of the letter-writer at the time of writing. But there is still an interesting way in which the letter case is different from AAS: a letter is (rarely) meant to ever convey agency independently of its writer, whereas AAS is. In that sense, AAS is closer to a book than to a letter. And it has been acknowledged that the original intention of the author of a book can come apart from the (successful) interpreted "intentions" of a book (at the time of reading). (Barthes, 2001) Even if we can map the abilities and agency of one onto the other, that doesn't mean it is explanatory to do so.

The AAS has a human equivalent, namely the memorised answers case from Chapter 3 (Section 2.iv). As a refresher, the problem with memorised answers was its lack of scope and sensitivity, not the mode of acquisition. Just because a human individual responds by using a phrase someone else thought of, a fact someone else discovered, or a formula someone else constructed, doesn't entail that that aspect of the individual's epistemic agency should get revoked - otherwise the endeavour to spread scientific understanding would be better accomplished by abolishing most schools, teachers

and textbooks, for they would merely serve to rob students of forming "their own" understanding with each preconceived idea they convey, each pre-discovered fact they share and each preformulated formula they relay. Just because we can causally trace each of AAS's actions to explanations outside of the agent (to a programming decision made by Ohce), doesn't entail that the postulation of an independent agent is a sterile endeavour.[225] If an entity can independently (in the sense of: without continuous outside help) produce certain appropriate abilities, and displays these abilities in a way that reveals a coherent and persisting subject, then attributions of understanding to that entity will be explanatorily powerful. It describes something about *that* entity. Nonetheless, this is only a small victory, because attributions of understanding in cases such as the AAS (or memorised answers) don't *only* describe something about *that* entity. They can also describe something about other entities, namely its programmers (or teachers). And these may perhaps be of equal relevance. After all, there is still a strong sense in which nearly "all of the relevant work" regarding responding appropriately was done by the programmer, even if all of the work was done ahead of time - and where the relevant work was done does make a difference in how we weigh up the relevance of the two postulates: that of the human epistemic agent and the artificial epistemic agent. So let us take a closer look at what that means.

The Longhand of Unique Origination

Even though the postulate of the epistemic stance can be explanatorily independent, there is still a sense in which it may not necessarily be the *best* explanation. The aspects of one explanatory entity may still justifiably regress to another if there is nothing *unique* about it. The reader will find some congruence between the uniqueness of an explanatory level (see Chapter 5) and the uniqueness of an independent entity. I will now offer the characterisation of regressability, constructed in explicit congruence with the characterisation of reducibility from Chapter 5 (the main difference being that you regress backwards rather than reduce downward):

> *Regressability*: An entity has regressable properties if those properties have a straightforward mapping-relation to their causal origin.

[225] It should be noted that it has also been argued that humans are also mere puppets of causal causes. For instance, Pereboom (2001) argues that there is no difference between being manipulated by someone and being determined by your nature and/or nurture since the latter is simply a more elaborate form of manipulation. If you accept both the manipulation argument and the regress problem, then even human understanding regresses. Aren't humans just a sum of their nature and nurture, both of which can be traced outside of themselves? So if causal origins entails superfluity, then even human epistemic agents are superfluous, because we can trace back any action down a causal history leading back to its teachers, birth, the process of evolution and ultimately, the big bang. (See also Coates & McKenna, 2015) For a more extensive argument against "the manipulation argument," see (Delarivière, 2015).

Let us look at the letter and AAS again through this lens. The letter is the most straightforward example of regressability. Whatever epistemic abilities a letter reveals (e.g. explaining a concept, or providing a proof, or outlining an experiment, etc) or agency we can infer from its words, it would be a mere subset of the abilities or agency attributable to the its author. If we were to postulate an additional epistemic agent, namely the letter, we would be attributing it with abilities, and explaining it with beliefs that would better suit our estimations of the author's future actions than those of the letter.[226] There is no benefit (nothing new) to seeing the letter as a separate agent, because one can be straightforwardly mapped, and therefore regressed, to (or derived from) the other. So in spite of the fact that the letter reveals these attributes, it cannot claim authorship over them.

Where Winograd and Flores hit on a legitimate worry is that programs like AAS can have their abilities or epistemic properties mapped to its creator almost as equally straightforward as those of letters onto its letter-writer. Ohce, the programmer, had to anticipate all of the relevant questions and answers that her program would be able to respond to, so that she could program in all the relevant canned responses to answer to the appropriate queries. Even though AAS can function independently of Ohce once it has been written (more so than the letter) and even though we can easily consider it as a separate entity and explain its actions through the epistemic stance without referring to the programmer (unlike the puppet, which is continuously dependent on the puppeteer), every ability we detect and every epistemic property we infer from them are as applicable to AAS as they are to Ohce. The epistemic stance targeting AAS at any time can be straightforwardly mapped (even overlapped) with the epistemic stance targeting Ohce at the time of programming. By contrast, most humans who memorise answers may grow into ever more unique understanders. Even if they still hold on to phrases, facts or formulas that they appropriated from their teachers, the regressability becomes ever less relevant in light of the whole epistemic agent.[227] But none of the successes of AAS will ever rise above those which Ohce explicitly anticipated and programmed in.

All properties we may ever detect in the world will have a causal origin, but not all properties have a *straightforward mapping relation* to their causal origins. If there was an explicit encoding of responses,

[226] If you prefer, you could swap "letter" for "email." An email is merely the digital counterpart of a letter. The medium is no longer a piece of paper, but a computer. But the text in the email does not reveal the agency of the computer as a separate entity any more than the letter revealed the agency of a piece of paper.

[227] Interestingly, the example relied on Ohce not changing much over time during the process of programming. If she did (e.g. because she keeps forgetting things she has already programmed in, or because her responses have become more precise than those she programmed in), then the combined epistemic abilities and properties we detect in AAS may not find a counterpart in Ohce at any particular point during her programming timeline. Then the epistemic agent of AAS would only regress to a conglomerate of Ohce's epistemic properties at different stages of her programming timeline. This also makes the epistemic stance targeting AAS more unique to AAS.

then any answer that the artificial system offers as a response, any ability it thereby displays, will have a straightforward mapping relation to the programmer who decided to put it in. But not all programs are mere explicit encoding of answers or even straightforward procedures, like AAS did.

Now consider a point made by Dennett (1997) about where originative credit is due by example of the computer chess-program Deep Blue. The chess program's behaviour is sufficiently complex and systematically purposeful to warrant an intentional stance. When Deep Blue beat world chess champion Garry Kasparov, it was "Deep Blue's sensitivity to those purposes and a cognitive capacity to recognize and exploit a subtle flaw in Kasparov's game that explain Deep Blue's success" (Dennett, 1997, p. 352) and not the designer's. The designers do, of course, play a necessary causal and intentional role in Deep Blue's cognitive capacity. However, congratulating the designers is much like congratulating the educators or parents of (or the process of natural selection leading to) Kasparov, his human opponent. (Dennett, 1997). Though it is certainly justified to admire Kasparov's teachers, parents or even process of natural selection, it doesn't take away Kasparov's credit as a world chess champion. It took Kasparov to be world chess champion. And it took Deep Blue to beat him. Deep Blue's programmers could not win against Kasparov. The sequence of winning moves was found by Deep Blue, not its programmers.[228] So it were Deep Blue's reasons and Deep Blue's intentions that are a proper target of using an intentional stance. The same can be said of AlphaGo's defeat of Lee Sedol in the game of Go. (Borowiec, 2016) AlphaGo's programmers may be credited with creating the program, but the winning moves were found by AlphaGo, not them.

The same claims can be made about more epistemically oriented artificial systems. For instance, there exists a Robot Scientist called Adam (King et al, 2004, 2009) that forms its own hypotheses, which it then proceeds to test. In its automated endeavours, Adam has rediscovered the role of genes of known function (King et al, 2004) and discovered new scientific knowledge about the genomics of Saccharomyces cerevisiae (King et al, 2009).

> "[S]ome of the mappings between genes and enzyme functions in S. cerevisiae that Adam has hypothesized and experimentally confirmed are certainly novel. Although this knowledge is modest, it is not trivial" (King et al, 2009, p. 53)

[228] The epistemic closure principle (which states that anything that can be derived from known information is also known) is widely criticised for exactly this reason. (Luper, 2020) We wouldn't want to credit Peano for every derivation in Peano arithmetic.

This discovery was not explicitly encoded in Adam, nor was the discovery made by its programmers. It was the hypothesizing and testing performed by Adam that resulted in the discovery. So when Adam expresses these results, it is expressing its own belief (a belief that is not, up until that point, to be found in any of the programmers at all) and it is its own (albeit very specific and limited) understanding it reveals, not that of its programmers.[229] There is no mapping relation of Adam to its programmers that would reveal its abilities (e.g. providing newly discovered knowledge) to be the abilities of its programmers - even if they are responsible for Adam's programming. This lack of a straightforward mapping relation becomes increasingly relevant if the artificial system's actions are not just a consequence of stimulus-response rules (like AAS), but a consequence of complex decision-making processes and learning, as in human individuals.[230] If artificial systems could only display the abilities of their programmers, then what would be the point of automated theorem provers, data analysers or robot scientists? What would be the point of creating any artificial system intended to aid in the process of science, beyond adding copies of the scientists and scientific discoveries we already have?

In sum, the regress problem can be found in those cases where the epistemic agent postulate is superfluous because its abilities and epistemic properties regress to causes outside of the agent. There are certainly cases where this can be claimed as a defeater of the artificial epistemic agent postulate (e.g. puppets) or a superior reading of the situation (e.g. AAS), but the claim that any postulate of artificial epistemic agency is superfluous takes the legitimate worry of a regressable epistemic stance, and unduly extends it to any case where there is a causal origin (e.g. teaching or programming), no matter how self-sufficient the entity is thereafter (which would furthermore lead, on pain of inconsistency to superfluity with human epistemic agents). If the epistemic stance lacks a straightforward mapping onto the epistemic stance of the creator (e.g. Deep Blue, Adam), then what makes the epistemic agent a uniquely powerful explanatory postulate would get lost in a regress-story. This entails there is no benefit, and even a disadvantage to changing the explanation from the program to its programmer. In short, "don't target the artificial epistemic agent, but the programmer" mistakes causal origin explanations for explanatory regression by failing to acknowledge cases of explanatory usefulness and uniqueness.

[229] Adam is undoubtedly very limited in its agency, and cannot be communicated with in the same way as human individuals, but it can form its own hypotheses, and it can reach its own results.

[230] We could once again craft the most extreme example: If we created an artificial expert by simulating a brain that was isomorphic to that of a human expert (like the Expert Planet, but with algorithms instead of citizens), would its abilities and agency be merely those of the programmer?

So it is not fair to say that the agency or abilities revealed by each program always regresses to those of its programmers in the same way that the agency revealed by letters, puppets or echos regress. Only if we can straightforwardly regress the acts of the (artificial) epistemic agent to something else, can we prefer the origin story over the epistemic agency story, because there is "nothing new". But the explanations we generate for the program and the programmer do not necessarily overlap such that it would be explanatorily beneficial to regress one to the other. Furthermore, the complexity of the mapping relation is a difference in degree, and not kind. It is true that this entails there is no ultimacy of authorship, but did we ever really expect there to be? People thank their teachers, their parents and their spouses, and not just to score social points.[231] We may be equally interested in the programmer as the program (we can do both), in the scientist and her verbatim references, but that doesn't mean one fully regresses to the other. We readily acknowledge that human achievements are not borne ex nihilo, so why must we hold artificial agents to a higher standard?

6.2 Artificial Epistemic Agents & The Reducibility Problem

But even if the artificial epistemic agent is explanatorily unique in comparison with the epistemic agency of its programmer(s), there is still cause for scepticism that targeting the artificial epistemic agent is the best explanation for the artificial system. Aren't programs just following the rules of their programming, so can't we just reduce the system's abilities to its design? A version of the *reducibility problem* rears its head for artificial systems as well as groups. This is the second conceptual hurdle. Fortunately, the assessment of the problem is very similar to what we saw in Chapter 5 for groups.

The Lovelace Objection & the Reducibility Problem

Earlier we read Lovelace's Objection of "nothing new" in the sense of "superfluous" and connected this to the claim that "everything can already be explained by its programmer." But there is another way to interpret "nothing new" or "superfluous," namely that the epistemic agent postulate is superfluous because there is a lower-level explanation which already explains the full situation (without postulating an additional virtual entity). If we can reduce the abilities and acts of agency to algorithmic procedures or components, then the epistemic stance is just a shorthand to talk about its design, its "programming," and therefore "not new." In short, we could claim that "everything can already be explained by its programming." This worry could have played a part behind Bringsjord's

[231] It is not irrelevant to note that humans tend to take pride in their contributions. Praise (as well as punishment) are used as incentives to cultivate the appropriate contributions because it is nice to be credited and painful to be unduly omitted. Therefore, being cheated out of credit is not just a metaphysical issue, but a moral one. As long as it makes no difference to artificial systems whether they get the credit or not (i.e. as long as they are not moral patients), there'll be an asymmetry in the warrant for credit-attribution. But even then, we must be aware of that reason for the asymmetry.

(2001) idea that a computer program can only to take it upon itself to originate something if it can reliable repeat an action for which it was not programmed and which cannot be explained even by the designer by appeal to the program's architecture, knowledge- base, and core functions. If a program relies on its knowledge-base, core functions and the architecture of its hardware, then any appeal to an artificial epistemic agent could be considered as superfluous (next to the appeal to its hardware, knowledge-base and core functions). This would be worrying for the explanatory uniqueness of the epistemic stance. Nevertheless, I will argue that it isn't. Not always.

Firstly, I must say that there is a legitimate point to be made about the reducibility of certain artificial systems' agency and abilities. Much like with epistemic group agency, we can find cases of artificial systems where the mapping relation makes the artificial epistemic agent a mere shorthand. As a reminder, here is how I characterised reducibility:

> *Reducibility*: An entity has reducible properties at a macro-level if those properties have a straightforward mapping-relation to any properties at a level below.

The macro-level here refers to the level of the epistemic agency or abilities, and the level below refers to the algorithmic (or hardware) level. If we can straightforwardly map the abilities or epistemic properties of the program as a whole to those of its algorithms, then the change of level is explanatorily superfluous. Then the epistemic stance is just a convenient shorthand to talk about its design, its "programming," and therefore "not new".

Consider APLI, a (hypothetical) Automated Propositional Logic Inferencer, which, starting from a few axioms and inference rules, attempts to generate a list of valid propositions (and their proof) in propositional logic. Even if we don't read its code, we can infer from its proofs which axioms and which inference-rules it deems to be valid. The epistemic stance would attribute these as beliefs. Nevertheless, we don't need to refer to a virtual postulate of epistemic agency. Any axiom or inference-rule it can be interpreted to believe, we will be able to find explicitly encoded somewhere in its code. Furthermore, any proposition it is able to prove can be reduced to a conjunction of smaller abilities (namely, the correct application of each inference-rule, along with some procedures for navigating the space of possibilities) which have also been explicitly coded. So both the abilities APLI displays and the epistemic agency (macro-systematicities) we can infer from them have a straightforward mapping relation to one of its algorithmic procedures (or an aggregate thereof). So one could be seen as a shorthand for the other. If the abilities or agency (macro-systematicities) of

the artificial system have a straightforward mapping relation to its algorithmic procedures (micro-systematicities), then those abilities or agential properties can be seen as a shorthand for those particular algorithmic procedures, exactly because we can systematically redescribe one explanation into the other without loss. While we can explain or predict APLI through the epistemic stance, if you can look at the relevant conjunction or cooperation of algorithmic sub-procedure, you can explain or predict APLI's responses a lot more readily. In short, we could reduce APLI's abilities and agency to its programming.

So there is a legitimate worry that the systematicities revealed by the epistemic stance do not posit additional benefits over the systematicities of the algorithms. This is the case when there is a straightforward mapping relation between the two, because that makes one a reducible shorthand for the other.[232] But that cases like APLI exist does not prove that there is a necessary dichotomy between humans and artificial systems. There is nothing inherent about artificial systems that says there must be a straightforward mapping relation. As I will show in the next section, there is not always such a straightforward sub-procedure as in APLI.

The Longhand of Computational Emergence

In its most extreme reading, the reducibility worry would have us assume that the explanatory power of the epistemic stance is only valid if the artificial epistemic agent can reveal itself *outside* of its programming or hardware (as seemed to be the case with Brinsjord). But if one tries to keep out reducibility worries by claiming that any mapping-relation between the epistemic agent and its implementation level would undermine that epistemic agency, then the epistemic stance would never be viable, even for humans, unless through magic. As we saw in Chapter 5, our brains supervene on the neurons of our brains (etc), so, given neuroscience (etc), there is a sense in which human epistemic agency is also superfluous and not "really new" either. But we also saw that what makes explanations invoking human epistemic agency distinct from neurophysiological explanations is that they focus on a systematicity that is uniquely tied to the macro-level. In similar vein, I will argue that the reducibility objection that artificial epistemic subjects would be superfluous because "it's just its programming" mistakes supervenience for explanatory reducibility by failing to acknowledge that macro-systematicities have explanatory power and can be explanatorily unique to the macro-level.

[232] I would like to repeat, however, that this doesn't entail we need to favour the explanation at the algorithmic level. We can be interested in the artificial system in isolation of how it is programmed. If it were to turn out that the beliefs of human individuals could be mapped to a certain region or systematic procedure in the brain, this does not entail we need to redescribe all of our ascriptions of belief to brain-procedures.

So far, we have only seen artificial systems that *do* have such a straightforward mapping relation. AAS or APLI, for instance, can be systematically described through the algorithmic level. When, for instance, we say "these are the exam-answers of AAS to question A, B and C" we can replace this with "these exam-answers are the composite of the answers of coded response A, coded response B, coded response C". Its abilities are distributed disjunctively, with the decision-procedure for which coded response will get printed being determined by another algorithm that parses the question and compares it with its list of encoded questions.[233] If we can reduce a program's actions to an addition, conjunction, disjunction or cooperation of algorithmic processes, like we could with AAS (or APLI), then we don't need to postulate an additional epistemic subject, because there's a way to systematically redescribe its abilities (and its agency) at the algorithmic level. Furthermore, if we can reduce a program's actions to an addition, conjunction, disjunction or cooperation of algorithmic processes, and those processes are directly encoded by a programmer, then there is a sense in which we can say that the artificial epistemic agent is superfluous because it both reduces and regresses. This is the case with AAS, where each response is explicitly coded by Ohce. But what if an artificial system is more than a mere aggregation of these processes?

I would like to roughly sketch some aspects of an artificial architecture (be it at the hardware or software level) to achieve the emerging effects we are talking about. I can't express it better than Forrest (1991)'s summary of the emergent computation approach:

> "Generally, we expect the emergent-computation approach (...) to have the following features: (1) no central authority to control the overall flow of computation, (2) autonomous agents that can communicate with some subset of the other agents directly, (3) global cooperation (...) that emerges as the result of many local interactions, (4) learning and adaptation replacing direct programmed control, and (5) the dynamic behavior of the system taking precedence over static data structures." (Forrest, 1991, p. 5)

As we can see, the focus here is on a distributed architecture which consists of a swarm of parallel subsystems interacting with one another (though not with complex information) in such a way to make up global effects. These global effects may enjoy something like an assembly bonus (or loss) effect. And, crucially, these global effects can't necessarily be mapped straightforwardly on the

[233] It is worthwhile to note that this is not quite true of its beliefs. Because all of its abilities to respond appropriately are explicitly encoded, the beliefs postulated by the epistemic stance can only be read "between the lines" or between the encoded responses, if you will.

algorithmic level. There is no central control system, no explicit encoding or formula that allows the artificial system to display the appropriate ability or respond appropriately. While the process at the algorithmic level may be as static and unchanging as the laws of nature, at a higher level, the system is flexible, can learn and adapt.

As an example, consider the connectionist approach. In the connectionist approach, algorithms are written that dictate the behaviour of artificial neural networks (see Chalmers, 1990; Smolensky, 1999). These artificial neural networks can then be used, for instance, to recognise speech or images. But there are no algorithmic procedures directly about speech or images, only about neural networks. The ability can only be detected at the macro-level.

> "[Connectionist models] have made familiar the notion that the level at which a system is algorithmic might fall well below the level at which the system carries semantic interpretation (Smolensky 1988)." (Chalmers, 1990, p. 658)

The ability of speech or image recognition as performed by artificial neural nets is a macro-systematicity. Furthermore, it is a macro-systematicity for which there exists no systematic mapping relation to its algorithmic level. While it was possible with AAS to redescribe its ability to answer questions appropriately as an aggregate of question-and-answers-related algorithmic stimulus-response procedures, there is no straightforward way to redescribean artificial neural network's ability to "recognise grandmother" as an aggregate of grandmother-recognising-related algorithms. Neural nets are famous for their reliability, but also for the difficulty in deciphering how exactly they process information. (Smolensky, 1999) This suggests a lack of systematicity in being able to map the macro-systematicity (the ability level) level onto its micro-systematicities (the computational level). If a systematicity at one level has no straightforward mapping relation to a systematicity at a lower level then we can't take advantage of the macro-systematicity at the level of micro-systematicity. So even if we can fully describe every artificial system at the algorithmic level, we would have to forego talking about certain artificial epistemic agents and their abilities if there is no systematic mapping relation between that artificial epistemic agency/abilities and the algorithms that implement them.

So not only can computers have abilities, but their abilities may be as difficult to account for in terms of aggregated algorithmic procedures as the abilities of human beings can be difficult to account for with a systematic neural network signals. There are certainly cases where the artificial epistemic agent and its abilities can be systematically mapped onto an aggregate of algorithmic procedures (e.g. AAS,

APLI), but the reducibility objection that artificial epistemic subjects would be superfluous because "it's just its programming" mistakes supervenience for explanatory reducibility by failing to acknowledge that macro-systematicities have explanatory power (e.g. in shorthands, where the macro-systematicites are convenient, even if they can systematically be redescribed as micro-systematicities) and can be explanatorily unique (e.g. in longhands, where the lack of systematic mapping-relation would force us to give up the macro-systematicities in favour of long winded redescriptions that apply only to particular cases at particular times). In short, this is as misguided a reading for certain artificial systems as it was for certain groups.

6.3 Artificial Epistemic Abilities & The Rigidity (or Informality) Problem

There is one more important concern that I haven't yet (directly) addressed. Even if artificial systems can be imbued with emergent, non-regressable agency and abilities, this does not entail that artificial systems can be imbued with the relevant abilities to match human scientists. What if artificial systems are just too rigid to match human intelligence? Such concerns popped up in the discussion regarding automated mathematics, and I will use it as the emblematic case for the *rigidity* (or *informality*) *problem* - the third conceptual hurdle. Fortunately, what we have seen about emergent understanding may allow us to counter this supposed limitation, or at the very least, revalue the strength of the criticism.[234]

The Epistemic Standing of Automated Mathematics

The use of computers in the practice of mathematics has only been a fairly recent phenomenon. And since mathematics has a reputation for being the formal, deductive science, it was hoped that its automation would quickly lead to impressive results. But, unfortunately, automated theorem provers have progressed slowly and produced little that is relevant to existing mathematical questions or problems. (Larson, 2005) Furthermore, the use of computers in mathematical research has provoked a fundamental discussion as to their epistemic standing as a method of mathematical inquiry. This peaked when the Four Colour Theorem (4CT) was proved by a huge amount of automated testing. (Swart, 1980) The discussion centred on three issues: (a) reliability, (b) surveyability or intelligibility and (c) capacity for understanding. Based on one or several of these, people have considered computer proofs to be: uninteresting or unsatisfying mathematics, a completely different sort of mathematics, or no mathematics at all. (MacKenzie, 1999; Vervloesem, 2007) This has sparked a

[234] The following assessments have been largely lifted from an earlier paper (Delarivière & Van Kerkhove, 2017), but the text has been reshaped to focus on artificial understanding and the rigidity problem.

debate about the differences or similarities between computer proof and its traditional human counterpart.

Tymoczko (1979) claimed that the lack of (a) reliability and (b) human surveyability of (lengthy) computer generated proofs entails that they depart from traditional proof, making the computer's accomplishments empirical and fallible. All we can do is put a degree of trust in a black-box. However, both computers and humans are subject to reliability and (sometimes) surveyability issues, making it hard to use these features to argue for a dichotomy between the two. While it is certainly true that computer-generated proofs can be overly long or complicated, they aren't always. And human generated proofs are not immune to length or complication either - narrowing expertise of ever more complex results see to that. (Geist et al, 2010) Furthermore, even when surveying is possible, the community accepts surveyed results without everyone partaking in the surveyability process, allowing human peer-reviewers to also function as the testimony of trustworthy black-boxes (Geist, Löwe & Van Kerkhove, 2010). Mathematics, it has been argued, remains as little (Burge, 1998) or as much (Swart, 1980) empirical when performed either by human or machine. Nonetheless, humans are considered as more trustworthy due to another quality they possess or supply. So even when surveying is possible, the question is what human surveyors supply that computers cannot.

It may perhaps seem odd that computers are not considered more mathematically able, given that mathematics has a reputation for being largely formal and rigid, much like computers. Under a traditional characterisation of mathematics, which focuses on detailed formal derivation according to rigid inference rules, mathematical practice would have been perfect for computers, which have a reputation for precise rule-following. If formal validity were the core of mathematical rigor, then no surveying but the formally rigorous kind would be necessary. This being closer to a computer's strong suit, their reliability alone would end the discussion. What computers are currently lacking, and what mathematicians seem to find most unsatisfying about them is something humans can deliver besides formal rigor. (MacKenzie, 1999; Avigad, 2008) What human surveyors (in the best cases) supply to warrant peer review and what provers supply that empower their proof is (c) understanding.

According to Rav (1999), this focus on understanding means the primary goal of mathematics is the development of mathematical meaning (conceptual interconnections or clarification) which cannot be derived from formal expressions, but instead requires active interpretation, an "irreducible semantic content" (Rav, 1999, p. 11). Currently, this lack of informal understanding often gets mentioned (MacKenzie, 1999) and is assumed to constitute a necessary difference, a dichotomy even,

and on the basis of such critiques, computers get pushed outside the realm of understanding and thus the locus of mathematics. But the critique is vague and little is done to explicate or investigate what this informal understanding might actually or preferably entail as well as when any of its characterizing criteria are met or left unsatisfied. Given our reading of understanding as a set of abilities, the critique is really just an objection about a lack of abilities. If certain programs are indeed bad at certain abilities, then their epistemic standing is as bad as their inherent lack of abilities. But if they aren't inherent, then the argument from informality falls flat of attacking artificial systems on the basis of it being an artificial system. So the strength of the critique boils down to an argument of inherent inability: namely informal abilities. And this is not a new argument.

The Argument of Informality

The most straightforward version of such an argument was labelled by Turing (1950/1985) as the Argument of Informality of Behaviour. He describes it thusly: "It is not possible to produce a set of rules purporting to describe what a [person] should do in every conceivable set of circumstances." (p. 65) This argument relies on the notion that programs need to be written by encoding the appropriate response for each set of circumstances, as was the case with AAS (where every query was linked up with a response) and even APLI (where there was a procedure for generating all the appropriate results). From this, it follows that every query that has not been encoded (i.e. formalised) will result in inaction or an error-message. A similar type of problem can be found in what Turing calls the Mathematical Objection (Turing, 1950/1985). The Mathematical Objection denies the possibility that computers could exhibit the characteristics of human thinking because they, unlike humans, are crippled by the limits of a formal system, such as the halting problem and Gödel's incompleteness problem. In that sense, they are also an argument from informality. I won't address the halting problem or Gödel's incompleteness problem directly, but I will address the larger claim: that artificial systems inherently lack certain abilities we value in humans because, unlike humans, artificial systems are always bound by their formal system. I shall subsume all of this in *the Argument of Informality*. Because it implies computers are too rigid, we can also call it *the rigidity problem* (which I have done partly for alliterative purposes with regress and reducibility). But before we can address the rigidity problem of artificial systems, it would do well to be able to pinpoint which abilities they are purported to lack and in what way they defy being captured by formal rules or algorithms.

To consider the Argument of Informality, let us work up from the traditional conception of mathematical practice. The traditional conception of mathematical practice takes mathematical proof (its key component) to be a matter of rigorous formal derivations aimed at justification and performed

in solitude. From this perspective, the corresponding characterisation of understanding mathematics would involve the ability to derive (all) consequences from well-delineated axioms according to strict inference rules. If this were what makes one understand mathematics, then the issue would really be settled by comparing the reliability of human and automated mathematicians to perform these inferences without error. This being closer to a computer's strong suit, their reliability alone would end the discussion. But a couple of things are wrong with this picture. First, the encoding of axioms and inference-rules won't do much to navigate the formal system. And even if one can find a procedure to navigate it fully (producing every theorem and exhausting every road to it), the process won't be efficient (the combinatorial explosion alone would yield it impossible in practice) and its search will be uninspired, blind to what makes a theorem or the route to it interesting. But there are further problems. If we conceive of the proving practice as formal derivation given the appropriate inference rules, then we could exhaust mathematical knowledge by fully navigating (and recording the routes within) a given formal system. But such a formal system is not a given. Instead, it is being shaped and reshaped by mathematicians according to their judgement. The same is true for the formation of concepts.

So we need a procedure for deriving interesting theorems (and doing so via interesting routes[235]) and we need a procedure for the judgement with which mathematicians improve or shape a formal system's axioms and inference-rules, as well as the concepts used. But how is this supposed to be accomplished? These judgements are not straightforward. Mathematicians sometimes choose between keeping a formal system with aspects which are un- or counter-intuitive, letting it shape new intuitions (e.g. axiom of choice, non-euclidean geometry), or keeping the intuition and adjusting the formal system. (Thompson, 1998) Furthermore, if one modifies the axioms of a formal system, one modifies the whole system, so whatever method of navigation or logic for discovery one uses will need to be accommodated to the space it navigates. Can we have a prefixed set of rules or algorithms that exhaust all the relevant axiom- and inference-modification as well as all interesting discoveries across all relevant formal systems? Can these judgements be captured by a formal meta-system, with the right meta-axioms and meta-inference rules? And if so, will it truly encompass the logic for mathematical discovery or should it itself be subject to further meta-considerations? And do these mete-considerations themselves need further meta-meta-inference rules? If so, what are the rules of the top-most meta-system (the complex rules that determine the results of all the systems)?

Perhaps one way to improve the discovery process would be to include the ability to recognise a good

[235] Finding interesting routes can be one of the reasons why mathematicians don't just prove, but reprove

thing when you stumble upon it. This no longer implies that the proposed process is determined to land on the interesting bits. Instead, it uses trial-and-error with various rules-of-thumb until it has found something it notes of interest. To accomplish this, we seem to need the meta-system to include both the ability to stumble with some wisdom (no trivial task) and an evaluation system that can gauge the interestingness of every derivation, axiom, concept or method it stumbles upon. Once again the question pops up: is there a universal standard of interestingness or is this open to change and development? As for the manner of stumbling, the same question pops up: are there universal rules-of-thumb or does this change with the space being explored and are these also rules-of-thumb subject to change according to one's (developing) interests? Furthermore, there is a high degree of interconnectedness between all these abilities or the rules that are supposed to capture them. How can this be accomplished within and across levels?

An even deeper problem lurks with this characterisation of the proving practice. So far, I have approached the problem through the traditional lens of mathematics being a formal system and the growth of mathematical knowledge being constituted by deriving theorems from these axioms. However, a group of 'mavericks', starting with Lakatos (1976), have challenged the view that formal derivation is the bastion of mathematics or its practice. Although formal proofs get valued for their theoretical rigor, the practice of formalisation is not only strenuous, but could also dramatically reduce a proof's intelligibility (Aberdein, 2006) and consequently become more prone to error than the usual more informal kind. (Harrison, 2008) That is not to say that mathematicians do not work with formal systems, but it is entirely misleading to reduce the proving practice to performing formal derivations. Instead, mathematicians produce proof outlines (Van Bendegem, 1989) which may (or may not) bear some direct relation to a full formal derivation, for example as an abbreviation or indication (Azzouni, 2004). In similar vein, instead of mathematicians using concepts according to their theoretical definition (that they may consciously endorse), their conduct indicates that what they really use are much vaguer and more fluid conceptions. The distinction has been noted as concept definition / concept image (Tall & Vinner, 1981) or manifest concept / operative concept (Tanswell, 2017). This bears importance because conceptualisation and proof formation are inextricably linked in the activity of mathematicians. Such things seem to indicate that, while human mathematicians may produce and work with formal systems, their thinking is not characterised by them. Mathematicians neither prove by navigating the search-space nor peer-review by checking proofs step by step for correct inference. What do they do then?

They rely on meaning, so we are told (e.g. by Rav, 1999). What could make up this meaning? Here is a

couple of broad strokes in answering that question: There is a great deal of recognition-processes going on in various ways, including identifying key elements or moves used in a proof and discerning the intentions, ideas, or approaches involved. What is also of importance is the recognition of patterns (in all aspects involved in the proving activity and at various levels of abstraction), which benefits from analogies (so to find and exploit similarities with other knowledge), intuitions (e.g. about the physical world - Lakoff and Núñez, 2000) or adapting methods from other areas (Cellucci, 2000). Other modes of reasoning can be used to exploit these, including visual reasoning or non-deductive inferences (Baker, 2015). Furthermore, the objects identified or patterns discerned are subject to various evaluations. For example, theorems can be important, beautiful, relevant (Larson, 2005), conjectures can be surprising or promising, questions interesting, concepts powerful, proofs explanatory, reliable, difficult or pedagogically successful (Aberdein, 2007) and so on. What's more, these evaluations are not made without connection to the previously mentioned processes of recognition, analogy, background intuitions and non-deductive reasoning. There is also a lot of trial-and-error involved here, including working with incomplete or ambiguously delineated information, relying on experience in one's judgement, making snap-judgements, learning to trust and when to trust in a systematic manner (Allo, Van Bendegem & Van Kerkhove, 2013). This last point is important to stress. No mathematician is an island. When we affirm that human mathematicians can survey or prove, it is also important to keep in mind that they are not, and need not be, able to do so ex nihilo. Some crucial aspects of their abilities (or results) may in fact rely on the presence of the larger practice (e.g. using other people's results, methods, judgements,...) or environment (e.g. use of calculator, pen and paper,...). It seems fair to say that the proving practice is driven by a large amount of knowledge and skills that are highly integrated with one another.

So rather than navigating *within* a pre-set rigorous system, the whole process seems more akin to bootstrapping itself *towards* a changeable formal system - starting from a general feel based on incomplete information and working oneself up, with various skills, towards formal rigor, and only up to the point where intelligibility is still possible. If humans use informal (vague, flexible or fallible) means to practice mathematics, then we have to consider the fact that these may play a functional, rather than peripheral role (if not in justification, then certainly in discovery). As such, these too have to be taken into account in automating an artificial mathematician. It won't do to exclude the "dirty" aspects of the kitchen, if these play an integral part in producing its meals. There will certainly be aspects of a kitchen that are simply unwelcome, but at this point, it may not always be clear which are valuable features and which are bugs.

The Level of Informality

The conclusion so far is that mathematics does indeed involve certain kinds of abilities, namely the informal ones. Doesn't this principally relegate mathematical understanding as outside of the realm of computers? If we contrast the informal practice with the formal approach in computers, it certainly makes their flaws less surprising. A computer's strong suit is its ability to handle brute-force calculations (as exploited, for example, in proving the 4CT) and computing according to well-delineated processes. Principal claims against automated reasoning and understanding, mathematical (Rav, 1999) or otherwise (Haugeland, 1979), do often invoke or imply the informal or non-formalizable nature of human reasoning. But is there sufficient reason to conclude that the realm of informal moves is unattainable for computers? At face value, it certainly seems so. After all, mathematical understanding is informal and open, whereas computers function rigidly formal. Therefore, informal computing sounds like a contradiction in terms. Nevertheless, I would like to argue why the possibility of *informal computation* should not be dismissed (yet).

The principal reason, I believe, why the notion of informal computation gets dismissed is because formalisation of intelligence is taken as a necessary condition for automation, or for artificial systems. To be sure, formalisation can be very useful to the enterprise of automated mathematics because it reduces mathematical "thought"-processes to something easy(-ish) to cast in an algorithm and to then automate: explicitly delineated definitions and inferences that aren't tarnished by the sloppy side-routes, ambiguous associations and dirty details of what went on in the human kitchen while cooking. But if human thought-processes are emergent systematicities, there is no guarantee that the emerging systematicity is a rigid pattern that can be shielded off from its sub-processes. Therefore, we may not be able to replicate thought-processes at the thought-level.

I would like to explain my point about computing emergent thought-processes with an analogy that was also used by Dennett (1986/1998) and Hofstadter (1982). If we would want to model the weather at the cloud-level, we would be forced to consider clouds as stable, well-delineated entities such that it can be safely ignored that clouds actually consist of molecules rushing out in different directions. Such a high-level approach has had success elsewhere. For example: the macroscopic properties of gas (e.g. volume, temperature, pressure) are stable enough to ignore the fact that they are actually composed of complex molecule-bumps at a lower level. But, unfortunately, the notion of "cloud" as well as "thunderstorm","cold fronts", "isobars" or "tradewinds" are not stable or well-delineated entities. So trying to model the weather at this level of abstraction may require too much simplification, too much to be lost in abstraction to allow the richness of weather to be captured by

an algorithm that concerns clouds. But this doesn't (at least in principle) determine meteorology to be a computational impossibility. There may be no rigid laws or patterns at the cloud-level that we can cast as algorithms, but there may be at the level below it. The computational level would then be sub-clouds, the level on which the clouds supervene. So if one were to succeed in capturing the molecule-level, the cloud-level would emerge with it. Furthermore, the systematicities that emerge with it would be equally non-rigid.

Now, what the previous exploration of mathematical practice seems to me to indicate is that we *won't* be able to collapse and ignore the lower levels that make mathematical thought possible in human beings. Not only is this formalisation incredibly difficult to accomplish, but it may also filter away nearly all the traces of the original meaning and discovery process – both of the result and the formalisation process. One can try to enrich an existing formalisation with a logic for discovery, but it is an open question whether there are justified laws of mathematical thought such that these can be replicated by an explicit algorithm without recourse to anything "sub-conscious." Disregarding what goes on in the kitchen below the step-by-step recipes of the chef would be ideal, but it may not prove to be possible.

Informal Computation

An alternative approach to automating mathematical thought-processes is to look for laws, not of thought itself, but of sub-cognitive processes (that collectively make up informal mathematical thought). Rather than automate the syntax of a well-delineated game (justified mathematical thinking), the focus is on automating the cognitive architecture (at some level of abstraction) of the game player or constructor. What is being automated then is not mathematical thought directly, but the architecture of the brain (at some level of abstraction) from which mathematical thought emerges. And what emerges is not necessarily rigid or directly formalisable (or at least not in every way). It is my contention that this substrate-level (i.e. a vast array of collaborating sub-cognitive processes) contributes more to mathematical abilities than was traditionally believed.

This is not to say that no mathematical thinking can (or should) be directly formalised for automation. Some of the (thought-)processes that implement abilities lend themselves quite well to direct formalisation for computation. For instance: brute-force calculation, doing integrals, etc. These processes deal with objects and manipulations that are well-delineated enough to allow capturing them as computations (usually with greater reliability than humans do). And to the extent that these formalised systems are used in, or useful for, mathematical practice it is worthwhile to automate them

directly. However, not all objects and manipulations that humans employ in their thinking seem to be so well delineated or rigid. And the assumption that a well-delineated system should suffice is betrayed by the realization that there are, in fact, large amounts of implicit information, vague intuitions and ambiguous associations that go into human mathematics. The difficulty of automated theorem proving seems to offer further evidence for this. Much like the objects of cloud dynamics (e.g. thunderstorms) can only emerge from the interactions of molecules, so some brainstorms (e.g. mathematical thinking) might only be able to emerge from sub-cognitive events. And if these sub-cognitive events do behave in a law-like manner, then that level will allow itself to be captured by an explicit algorithm.

This line of reasoning might seem to strongly suggest a neuro-physical approach (i.e. simulating the brain) to achieve anything like artificial mathematicians. But my claim is not that there are only two options: either to formalise thought or to simulate the brain. It is just that I believe, like Hofstadter (1982), that any AI model "has to converge to an architecture that at some level of abstraction (so not necessarily at the hardware level) is "isomorphic" to brain architecture, at some level of abstraction" (p. 15), and this is not necessarily at the neuron level. This level could be anywhere, but it seems clear from both the limited successes of automated mathematics and from how the traditional view of mathematical practice has been criticised that this level will be considerably lower than that of thought - otherwise "laws of thought" or their corresponding formal system would suffice to capture mathematical thinking.

Bearing in mind the distinction between the level at which objects of thought can be identified and the level at which computable laws exist, remember Forrest (1991)'s summary of emergent computation:

> "Generally, we expect the emergent-computation approach (...) to have the following features: (1) no central authority to control the overall flow of computation, (2) autonomous agents that can communicate with some subset of the other agents directly, (3) global cooperation (...) that emerges as the result of many local interactions, (4) learning and adaptation replacing direct programmed control, and (5) the dynamic behavior of the system taking precedence over static data structures." (Forrest, 1991, p. 5)

This seems amiably suited to achieve a less rigid way to implement the processes behind informal abilities. The focus is on a distributed architecture which consists of a swarm of parallel subsystems interacting with one another (though not with complex information) in such a way to make up global effects. It is these global effects which we would call "thought," and they are the result of the cooperating subsystems, not a rigid rule or formula. While these subsystems may be as static and unchanging as the laws of nature, it is the global level where the system learns and adapts.[236] This architecture allows the possibility that "pieces of evidence can add up in a self-reinforcing way, so as to bring about the locking-in of an hypothesis that no one of the pieces of evidence could on its own justify." (Hofstadter, 1982, p. 14) For instance: Mitchell and Hofstadter's (1990) Copycat model is a case that satisfies the conditions of emergent computing. Copycat attempts to implement cognitively plausible high-level (and non-algorithmic) processes for anagram-solving by means of interactions between a number of low-level (but algorithmic) agents. And Chalmers (1990) has said of the model that it "is able to come up with "insights" that are similar in kind to those of a mathematician" (p. 659).

An emergent artificial system may come with the price of being fallible, but also with the benefit of possible continuous self-correction and improvement. The notion of decidability and its subsequent problems (as in the Mathematical Objection) are no longer fitting, because they apply to the computational level, and this is not the level at which mathematical decisions get made. The system does not simply compute until it has terminated upon the solution (or goes on ad infinitum or ad error). Instead, the sub-cognitive processes will keep on going with "the relatively mindless and inefficient making and unmaking of many partial pathways or solutions, until the system settles down after a while not on the (predesignated or predesignatable) "right" solution, but only with whatever "solution" or "solutions" "feel right" to the system." (Dennett, 1986/1998, p. 227) Or because another problem, idea or peculiarity draws it away from the previous one, as it can do with human mathematicians as well.

It may once have seemed that (research) mathematics would be one of the easiest of cognitive processes to automate, but it turns out it may be one of the most difficult. The objects and manipulations of mathematical thinking in practice are not rigid, simple and well-delineated enough to always allow capturing them in formalisations (which hush away so much of the mathematical "thinking" and discovery-process), so the automation of such a formal system may only lead to very limited results. Furthermore, considering how difficult it is to formalise all of mathematics and that it

[236] Learning could itself be seen as a dimension of informality, because a learning system is not bound by a preset formalisation (even if its learning mechanism might be).

doesn't seem that high upon the list of a mathematician's concerns, it seems important to try to automate something closer to the informal mathematics as it is practiced. Since mathematical thought-processes emerge from the architecture of the brain and since they furthermore appear to defy formalisation to such an extent, it may be sub-cognitive processes on which we will need to focus if we want to create an artificial mathematician. Then we no longer (necessarily) speak about a logic of discovery, but simply a process of discovery. Not a process designed to consistently and exhaustively run through mathematical truths, but a process that thinks - makes assumptions, recognises patterns, tries out methods, questions its own rigor - and thereby climbs up to what is mathematically convincing.

Whether it is realistic to expect this approach to be successful in automating mathematical understanding is far from proven, but the argument for its in principle impossibility has lost its strength. If understanding a certain scientific topic is composed of informal abilities (meaning they cannot be exhausted by a formalised rule or set of rules), then any artificial system that relies on a directly formalised process will lack in quality (sensitivity and scope in particular). But this alone does not doom artificial systems to remain outside of the practice of science, since the computational level may fall well below the level of its abilities. There is hope for emergent computation in the form of a decentralised system composed of dynamically cooperating sub-processes forming global learning and flexible abilities. Only if there is no conceivable level from which the ability could emerge would it exclude the possibility of automation, but it may, with it, raise questions about how humans perform their "magic", if they do.

6.4 Artificial Understanding

Let us take stock of what we have learned and what this means for artificial understanding. I shall quickly run through the conceptual hurdles we have overcome and end the chapter with giving examples of how the road to artificial mathematicians is being trodden in the wild.

Overcoming the Conceptual Hurdles

It was quite easy to establish that artificial systems can display abilities and reveal a coherent persisting agent. But, as we saw in this chapter, there were still three reservations about attributing understanding to the artificial system (barring a lack of abilities or persisting coherence), namely: the regress problem, the reducibility problem and the rigidity (or informality) problem. These problems formed conceptual hurdles that we had to overcome to justify the conceptual possibility of artificial understanding. The first two conceptual hurdles were based on the Lovelace Objection. This involved

both the regress and reducibility problem. Overcoming these hurdles involved having an answer to the question: Why doesn't the understanding of an artificial system automatically regress to its programmer or reduce to its programming? And the last conceptual hurdle involved an argument of inability, namely informal abilities. Overcoming this hurdle involved having to answer the question: Why aren't artificial systems too rigid to display the full scope and sensitivity of (informal) abilities we find in human beings? Let us reconsider each hurdle in their own turn.

First, consider the problem of regressability: If the abilities of the artificial system have a straightforward mapping relation to those of its programmer, then the attribution of those abilities to the artificial system can be seen as a shorthand for the attribution to that programmer without loss. In short, we can regress the abilities. This irregressability also extends into the epistemic subject as a whole. If the epistemic agency of the artificial system has a straightforward mapping relation to that of its programmer, then that epistemic agency can be seen as a shorthand for that of its programmer (at the time of programming) without loss. In short, we can regress the epistemic agency, making the postulate of epistemic agency superfluous. Nevertheless, that doesn't entail we *need* to regress, as we can be interested in the artificial system in isolation of its programmer(s). Furthermore, if there is no such straightforward mapping relation (e.g. because it was written via trial and error, or by several people dynamically), then the explanatory benefit from focusing on the artificial system's abilities or epistemic agency cannot be referred to by focusing on the programmer(s) (even if the programmer's abilities or agency may account for every line of code). This was clearest in the case of AlphaGo (or Adam), where the abilities of the programmers may account for the existence of the program, but AlphaGo's defeat of Lee Sedol could not have been achieved by any of its programmers or a straightforward aggregation of their individual abilities.

Second, consider the problem of reducibility: If the abilities of the artificial system have a straightforward mapping relation to its algorithmic procedures, then those abilities can be seen as a shorthand for particular algorithmic procedures, because explanations of abilities can be redescribed with those particular algorithmic procedures without loss. In short, we could reduce the abilities. This irreducibility also extends into the epistemic subject as a whole. If the epistemic agency (macro-systematicities) of the artificial system revealed by its acts (including its abilities) has a straightforward mapping relation to its algorithmic procedures (micro-systematicities), then that epistemic agency can be seen as a shorthand for particular algorithmic procedures, because explanations of its agency can be redescribed with those particular algorithmic procedures without loss. In short, we could reduce the epistemic agency, making the postulate of epistemic agency superfluous. Nevertheless, that

doesn't entail we *need* to, as we can be interested in the artificial system in isolation of how it is programmed. If it were to turn out that the beliefs of human individuals could be mapped to a certain region or systematic procedure in the brain, this does not entail we need to redescribe all of our ascriptions of belief as really being references to brain-procedures. Furthermore, if there is an assembly bonus (or loss), there is no such straightforward mapping relation and then the (macro-)ability or artificial epistemic agent cannot be referred to by staying at the level of the algorithms (even if a full algorithmic description would include it). This was clearest in the case of neural nets, where the algorithms describe the behaviour of the neural nodes, not the task (or ability) they are performing, which has no straightforward mapping relation to the algorithmic procedures of the nodes.

Lastly, consider the rigidity (or informality) problem. If understanding a certain scientific topic is composed of informal abilities (meaning they cannot be exhausted by a formalised rule or set of rules), then any artificial system that relies on a formalised process will lack in quality (sensitivity and scope in particular). But this alone does not doom artificial systems to remain outside of the practice of science, since the computational level may fall well below the level of its abilities. There is hope for emergent computation in the form of a decentralised system composed of dynamically cooperating sub-processes forming global learning and flexible abilities. This entails that artificial understanding is attained if an artificial system displays abilities, from which we can successfully postulate an explanatory or predictive epistemic agent, where there's no straightforward mapping relation between the abilities or agency one the one hand, and either the program that implements it or the programmer(s) that wrote it on the other hand. For abilities that do not allow themselves to be easily formalised, automation may need to focus not on the level of cognition, but the level of the cognitive substrate, or any level in between.

So even though artificial understanding has several conceptual hurdles to overcome, I have shown that none of them banish artificial systems from the possibility of understanding. At least not in principle. This is, of course, a different question to whether any implementations of artificial understanding will ever be in practice. Nevertheless, there is some cause for (cautious) optimism. To end, I would like to give an example that may aid that optimism.

Artificial Understanding in the Wild (HR & Leo-III)

When it comes to bringing up interesting cases from "in the wild," it seems appropriate to keep our focus on automated mathematics. In the field of automated mathematics, we can find a small group

of people who are attempting to automate mathematical discovery and concept formation, letting computers explore (Hales, 2008). I shall briefly refer to just two projects that caught my eye.

The first concerns the *HR-system* and its extensions. For the HR-system, inspiration was taken directly from the philosophy of mathematical practice. HR forms concepts and conjectures. While it does rely on strict production rules for its concept formation, the interplay with conjecture-making (which includes evaluations of interestingness as well as parsimony, novelty and surprisingness) and theorem-proving (which it outsources to OTTER, another automated theorem prover) make HR promising. (Colton, Bundy & Wash, 1999) This is doubly true for the extended HR-L, a multi-agent system which models interaction between different copies of HR running concurrently (each gauging interestingness differently). This has been said to lead to "greater creativity in the system as a whole" (Colton, Bundy & Wash, 2000, p. 16). Pease (2007) presents HR-L as a computational reading of Lakatos's theory of mathematical discovery and justification, learning from his suggestions of ways in which concepts, conjectures and proofs gradually evolve via interactions between mathematicians. Furthermore, inspired by Lakoff and Núñez's (2000) theory of embodied mathematics, Pease et al (2009) explore an analogical process to construct complex mathematical ideas (including both theory and axioms) via a combination of innate arithmetic and grounding metaphors. There is another extension of HR, called HR-V which uses pattern recognition on analogous visual representation for concept formation in number theory. (Pease et al, 2010) Though it can't as of yet generate these diagrams (and is thus much reliant on human individuals), I consider its use of visual pattern recognition for concept formation as progress in one of the crucial aspects of intelligence underlying mathematical understanding.

Benzmüller et al (1999, 2001) also seem keen to take many of the previously mentioned ideas to heart, aiming to emulate the flexible problem solving behaviour of human mathematicians in an agent based reasoning approach. They have proposed a multi-agent architecture for proof planning consisting of a society of specialised reasoning agents, each of which has a different strategy and work in both competition and cooperation with one another. A resource management technique is used to periodically evaluate an agent's progress (and thus how much resources to be allocated) and allow restricted communication amongst them about successful and interesting unsuccessful proof attempts or partial proofs, from which other agents can learn using a reinforcement learning approach. Their most recent agent-based project in that same line is called Leo-III and it is a multi-agent software where each agent functions as an autonomous specialist employed for some aspects of proof search. The underlying architecture is designed as a blackboard that agents can

collaboratively use in their process of finding a proof, having the work divided and auctioned off. (Steen, Wisniewski & Benzmüller, 2016)

Neither the HR-systems nor Leo-III come with previously encoded (or even expected) results, and the process of navigation of the search-space is too complex to reduce or regress its results to a straightforward aggregate of algorithms or the agency and understanding of its programmers. These systems still have fairly traditional features (most notably in that their results are very much bound to certain limits of a particular formal system), but their increased abilities seem to be due to their attention to embracing the flexible trial-and-error process of discovery of an informal mathematical practice, and I applaud them for that very reason. [237]

The progress regarding the quest for artificial understanding has been an impressive, but slow one. The biggest problem with even the most advanced artificial systems is one of scope and sensitivity. Although they can often outperform humans in some precise areas, I have as of yet no knowledge of systems that show a wide scope of abilities, such as producing outlines, analogies, interactive explanations, use of the relevant information at the epistemically relevant time, etc. Many of even the best examples of artificial understanding find their abilities in brute-force calculation techniques, or in (mostly) less advanced understanding than we would find in human experts. Nevertheless, as I have shown in this chapter, the principal objections that supported the claim that artificial understanding is inherently out of reach do not stand up to scrutiny. Given this, it is my contention that we have no reason to suspect that the possible advancements of automating scientists are soon to be exhausted. Achieving human-like abilities is a difficult endeavour, but maybe we shouldn't (yet) exclude the possibility that computers could play a much more meaningful role in the scientific practices - not just as a method of inquiry, but as fellow inquirers, as artificial understanders.

In Sum

Could artificial systems ever be considered as subjects with understanding? It was quite uncontroversial to establish that it is possible for artificial systems to display epistemic abilities such that the epistemic stance is explanatory or predictive. But, as we saw in this chapter, there were three further worries about attributing understanding to artificial systems that seemed to be based on the very nature of those systems. Those worries were the regress problem, the reducibility problem and the rigidity (or informality) problem. These problems form three conceptual hurdles that we had to

[237] Furthermore, in both HR-L and Leo-III, the overall system and its composing subsystems may benefit from the epistemic stance, entailing that we have a clear case of artificial collective understanding.

overcome to justify the conceptual possibility of artificial understanding. The first two conceptual hurdles were based on the Lovelace Objection, which claimed that artificial systems cannot originate anything new outside of what we tell them to do. This involved both the regress and reducibility problem. The third conceptual hurdle is based on assumptions about the nature of both human individuals and artificial systems.

The regress problem is the worry or criticism that the epistemic agent postulate is superfluous because its abilities and epistemic properties regress to causes outside of the agent. There are certainly cases where this regress can be claimed as a defeater of the artificial epistemic agent (e.g. puppets, letters) or a superior reading of the situation (e.g. AAS) because nothing explanatory is lost in the regress. If we can straightforwardly map the acts or epistemic properties of the (artificial) epistemic agent to its origin(s), then that epistemic agent postulate provides nothing explanatorily new or distinct over the regression to that of its origin(s). But the claim that any postulate of artificial epistemic agency must regress takes the legitimate worry of a superfluous epistemic stance, and unduly extends it to any case where there is a causal origin (e.g. teaching or programming), no matter how self-sufficient or distinct the entity is thereafter.[238] If the epistemic stance targeting the artificial system lacks a straightforward mapping onto the epistemic stance of its author(s) (e.g. Deep Blue, Adam), then what makes the epistemic agent a uniquely powerful explanatory postulate would get lost in a regress-story. This means there is no benefit, and even a disadvantage, to changing the explanation from the program to its programmer. So it is not fair to say that the agency or abilities revealed by each program always regresses to those of its programmers in the same way that the agency revealed by letters or puppets regress. Furthermore, the complexity of the mapping relation is a difference in degree, and not kind. It is true that this entails there is no ultimacy of authorship, but we readily acknowledge that human achievements are not borne ex nihilo, so why must artificial agents be held to a higher standard?

The reducibility problem is the worry or criticism that the epistemic agent postulate is superfluous because its abilities and epistemic properties reduce to its algorithmic procedures. But the abilities of artificial systems may be as difficult to account for in terms of aggregated algorithmic procedures as the abilities of human beings can be difficult to account for in terms of systematic neural network signals. There are certainly cases where the artificial epistemic agent and its abilities can be systematically mapped onto an aggregate of algorithmic procedures (e.g. AAS, APLI), thus making the additional epistemic agent an explanatorily superfluous postulate, but the reducibility objection that

[238] This would furthermore lead, on pain of inconsistency, to the superfluity of epistemic agency even in human individuals, given a nature (i.e. evolution) or nurture (i.e. teachers) story.

artificial epistemic subjects would be superfluous because “it's just its programming" mistakes supervenience for explanatory reducibility by failing to acknowledge that macro-systematicities have explanatory power (e.g. in shorthands, where the macro-systematicites are convenient, even if they can systematically be redescribed as micro-systematicities) and can be explanatorily unique (e.g. in longhands, where the lack of systematic mapping-relation would force us to give up the macro-systematicities in favour of long winded redescriptions that apply only to particular events at particular times). In short, this is as misguided a reading for certain artificial systems as it was for certain groups.

Lastly, the rigidity problem is the worry or criticism that artificial systems are inherently limited in their abilities due to being a formal system. If understanding a certain scientific topic is composed of informal abilities (meaning they cannot be exhausted by a finite formalised rule or set of rules), then any artificial system that relies on a formalised process will lack in quality (sensitivity and scope in particular). This seems plausible, given that many of even the best candidate examples for artificial understanding find their abilities in brute-force calculation techniques, or in (mostly) less advanced understanding than we would find in human experts. Nevertheless, as I have shown in this chapter, the objections that claim understanding is inherently out of reach for artificial systems does not stand up to scrutiny. The rigidity problem mistakenly assumes that the level of computation must align with the level of abilities, but the computational level may fall well below that level. As we saw in discussing reducibility and emergence, assembly bonuses may emerge from limited low level systematicities. Now, since epistemic thought-processes emerge from the architecture of the brain, and they furthermore appear to defy formalisation to such an extent, it may be that we will need to focus on formalising the appropriate sub-cognitive processes if we want to create artificial mathematicians. Here, the level of computation is lower than that of the emerging mathematical abilities. There is hope for emergent computation in the form of a decentralised system composed of dynamically cooperating sub-processes forming global learning and flexible abilities. Given this, it is my contention that we have no reason to suspect that the possible advancements of automating scientists are soon to be exhausted. Achieving human-like abilities is a difficult endeavour, to be sure, and it may even prove impossible in practice, but we shouldn’t (yet) exclude the possibility that computers could play a much more meaningful role in the scientific practices - not just as a method of inquiry, but as fellow inquirers, as artificial understanders.

DISSERTATION

Conclusion

Given that the value of understanding is hard to deny (because understanding is a valued aim and trait in many activities and disciplines), and that the value of its mark is no longer denied (since the concept of understanding has dissociated itself from its psychological dimension, as well as distinguished itself from the concepts of explanation and knowledge), we were in need of a conceptual characterisation that is explanatory as well as philosophically coherent and consistent, and which furthermore allows us to explain who does and does not understand, as well as why or why not.

I have, in this dissertation, set up my conceptual characterisation of understanding and the understanding subject with a big picture approach. I have shown that this approach can reveal a coherent picture of understanding and its subject. One which can tie together several insights from a variety of fields regarding the many aspects of understanding, and deal with many of the problems and objections revealed by them, without having to deal with them through their respective characterisations or insights in isolation of others. I will briefly recap my account here.

Characterising Understanding

In the first chapter, I focused on *the mark of understanding,* namely which systematic trait we find so philosophically or epistemically valuable about understanding and thus necessary for its attribution. This mark of understanding needed a philosophically coherent and explanatory characterisation that can be applied consistently to various human subjects (and beyond), and across various objects with varying degrees of (contextual) quality. Furthermore, it needed to allow us to deal with the known philosophical problems of marks, and address possible counter-examples.

I have argued that understanding-attributions always boil down to a particular set of *appropriate abilities* (of a subject), composed of acts (salient to the object for a certain context), and that this is the most coherent and useful conceptualisation of "understanding." There are many benefits to an ability approach: we side-line (without discarding) the mistrusted role of feelings. We avoid some of the problems that plagued mental state-based approaches, such as its looking for a mark in an empirically unobservable realm (we cannot discern anyone else's mental states, and even struggle adequately characterising our own), its explanatory redundancy (it is not the mental states themselves that are empirically accessible or epistemically valuable to us, so we both detect and judge mental states by the abilities, and not vice versa) and its required infinite encoding (every component of

understanding would need to be encoded as a state, quickly leading to an infinite regress). By contrast, ability-based approaches do entail considering what lies *beyond* observable acts (through explanatory estimations based on observed acts, conceptualised modally as counterfactual acts), but not what lies *behind* them (in an empirically unobservable realm), as mental state-based accounts presumed were necessary. Furthermore, the concept of implicit understanding is given more room to flourish (because epistemic abilities can be valued even without the subject being characterised as "aware" of them), and the problem of implicit chauvinism is given less room to flourish (it is harder to substantiate that a particular gender, ethnicity or even species lacks understanding if one has to mark a valuable difference in performance rather than in physical or presumed mental constitution). And lastly, some of the useful concepts associated with understanding that do not obviously match with an act-based approach (e.g. beliefs and meaning) can not only keep their explanatory power, but are even more firmly rooted as instrumental concepts derived from acts.

Having established abilities as the mark of understanding, I briefly considered some candidate kinds of abilities offered by the literature as the appropriate one(s). I then indicated that I will consider none of them as a necessary or sufficient condition for understanding, but instead as what composes the scope of understanding. This dissolves the need for unwieldy conditions (e.g. anti-luck conditions) and fits with the idea that understanding is not binary, but comes in levels or degrees. To explain that, we needed to conceptualise the quality of understanding.

Characterising its Quality

In the second chapter, I conceptualised the *dimensions* and *degrees* of *quality* in understanding, offered up a contextual approach to specifying what is salient, and specified some of the problems, opportunities and virtues in evaluating understanding under that approach. Even though most authors acknowledge the degrees of understanding, few of them address them as explicitly as was done here.

We saw four dimensions of quality, three of which were composed of two parameters, one which widens it and another which deepens it (see summary below). Unfortunately, and quite unsurprisingly, no agreed single universal standard can clarify all attributions of understanding within these dimensions. We found contextual variance not only in thresholds, but in what is considered salient in the first place. Therefore, I also offered up a contextual approach the dimensions and parameters, allowing each of them to vary what is appropriate for a particular context of attribution, while leaving the justification of what is an appropriate context of attribution to another discussion (see summary below).

Dimension	*Parameters*	*Contextual light*
Scope of abilities	*Scope* the amount of variety in abilities salient to understanding an object X.	*Scope (or domain) weights* which abilities connected to X are deemed salient, and to what extent.
Sensitivity of an ability	*Situational responsiveness* amount of appropriate changes in performance to changes in the object-situation, e.g. responding to what-ifs.	*Situational or what-if weights* which variations in object-situations, along with their appropriate reactions, are relevant, and to what extent.
	Accuracy degree of precision in performance, e.g. number of decimal points.	*Accuracy weights* which types of accuracy are appropriate (when there are degrees to success) and to what extent.
Stability of an act	*Range* degree of presence in (counter-) factual circumstances.	*Range or deployment weights* which types of (counter)factual circumstances, where the same epistemic subject acts appropriately, are salient and to what extent.
	Robustness degree of presence after (counter-) factual circumstances.	*Robustness or rationality weights* which types of circumstances, after which the subject needs to continue to act appropriately, are salient and to what extent.
System Efficiency of a subject	*Economy* the appropriate act uses a minimum of saliently allowable resources.	*Economy or resource weights* which particular resources (incl. events) one is willing to allow to be used to consider the abilities achieved, and to what extent..
	Potential the appropriate act obtains with the addition of a minimum of salient resources or events.	*Potential weights* which particular resources and circumstances one is willing to consider in order to assess the abilities achieved with it, and to what extent.

It is my contention that most attributions of understanding will boil down to a claim about the degree within these dimensions. Even if these parameters are imperfect in conceptualising an "ideal" or quantitative assessments of the quality of understanding, they are fruitful in diagnosing the strengths,

weaknesses, kinds and differences in quality as well as the problems or opportunities in evaluation (e.g. kludges, indirect vs direct evidence, kinds of understanding complete understanding). This was attested by how well my account fared in addressing tricky or misleading attributions and proposed counterexamples to the ability account in Chapter 3.

Characterising its Subject

Up until recently, it was often assumed that the targets of understanding attributions are (or should) always be human individuals, so the subject with understanding has not been considered in the epistemology literature with equal care as the mark of understanding has. Nevertheless, there are many cases that challenge that assumption, and we need a way to conceptualise this with consistency and without an anthropocentric bias. Many of the abilities we find in science and everyday life today are more and more frequently displayed by entities composed of more than (just) human individuals. Abilities are displayed with the use of resources (e.g. pen and paper to work out a proof, a calendar to remember a step in an elaborate experiment, an interactive theorem prover to discover new mathematical proof, any number of apps, etc), constituted by a group of individuals (e.g. a disjunctive division of answers in a pub-quiz teams, a conjunctive or additive pool of data in a research centre, a compensatory average of a crowd, a cooperative division of labour in a lab-experiment or a dynamical interaction between a couple) or even constituted by artificial systems by themselves (e.g. navigation systems, automated programs that deduce physical laws or mathematical proofs, or software that predicts traffic flow, rainfall or storms). What we need is a *mark of epistemic subjecthood* that would help us target a relatively cohesive, coherent and persisting entity such that attributions of epistemic properties (e.g. understanding, beliefs, etc) would be explanatory or predictive. In Chapter 4, I have focused on this mark.

As a mark of epistemic subjecthood, I have defended the interpretationist approach, and more particularly the *epistemic stance* (the intentional stance with an epistemic focus). The epistemic stance is the strategy of interpreting behaviour by treating it as if it were governed by beliefs, epistemic aims (i.e. the kind of results that an epistemic practice values), and epistemic tactics (i.e. any serious systematic attempt to get closer to an epistemic result), as well as any other intentions that play a supporting role in epistemic agency. It is instrumental in that the sole justification for interpreting an entity as an epistemic agent is the explanatory and predictive success of that interpretation. Nevertheless, the epistemic stance allows us to conceptualise the inner workings (e.g. beliefs, aims, tactics) from an act-based perspective, and demarcate the agent on the basis of all this (by looking for the realising base for the postulated epistemic agent).

Epistemic agency can be detected thanks to macro-systematicity. Systematicity is a pattern (i.e. the epistemic agent) that a theory (the epistemic stance) can predict or explain. The virtual postulate it reveals is *holistic* (as opposed to atomistic), meaning that components can only predict or explain the behaviour as a whole. Nevertheless, only if the components (i.e., beliefs, aims and tactics) that we ascribe to an entity are relatively *coherent*, can they be explanatory (e.g. if a subject is ascribed with contradictory beliefs, we will generate contradictory explanations or predictions). The epistemic stance thereby makes sure that the targeted virtual entity is tied together with (relative) coherence. And because the systematicity of the epistemic stance is at a higher (macro) level, it makes no dictates on implementation (outside of realising the systematicity). It does not rely on there being a direct correspondence between our ascriptions of beliefs, aims and tactics and some structure in the brain. The components of the epistemic stance are virtual, not physical. Nevertheless, a single epistemic stance will only be explanatory if there is physical *cohesion*. Only if parts of the world interact with one another would it make sense that the realising base (for the epistemic agent) will involve all those parts. The epistemic stance thereby makes sure that the targeted entity is tied together with physical cohesion.

So could coupled systems, groups or artificial systems ever be a successful target of the epistemic stance? This was the focus of Chapters 4, 5 and 6 respectively. Of course, there was no guarantee that the abilities of entities beyond human individuals necessarily lead to the success of an epistemic stance. There could be failures of the epistemic stance due to a lack of cohesion (e.g. Otto doesn't adhere to the information he chose to write down if he doesn't take his notebook with him, the exam answers in the Composite Class were produced without discussion, and two software programs on the same computer may determine their results in isolation) or lack of coherence (e.g. Lenny behaved erratically because of her flawed system, the Summative Class suffered from the discursive dilemma, and software can produce contradicting results) or persistence (Otto can lose his notebook, Lenny's system can be overhauled by outside influence, computers can reboot with a clean memory, and even human beings can die or suffer an accident). But if the epistemic stance is successful (e.g. Otto, Olaf, Coqto, Jointly Committed Class, Expert Planet, CERN, Google Maps, Adam, Deep Blue etc) then taking advantage of its explanatory power is not only warranted and fruitful, but also consistent with our best conceptualisations of (human) individuals.

Overcoming the Reducibility, Regress and Rigidity Problem

As we saw, there were still three further worries about attributing understanding that didn't (at first) seem to be a problem for human individuals. Those worries were the reducibility problem, the regress

problem and the rigidity (or informality) problem and they formed three further conceptual hurdles that we had to overcome to justify the conceptual possibility of understanding beyond human individuals.

The *reducibility problem* was the worry or criticism that the epistemic agent postulate is *superfluous* because its abilities and epistemic properties *reduce* to an explanation *without the agent*. This was a worry for extended epistemic agents (i.e. we can focus on the human individual embedded in its environment instead), collective epistemic agents (i.e. we can focus on the members of the group instead) and artificial epistemic agents (i.e. we can focus on its algorithmic procedures instead). But, due to the assembly bonus effect, the abilities of extended, collective or artificial entities may be as difficult to account for in terms of embedded human actions, aggregated member contributions, or algorithmic procedures as it is difficult to account for the abilities of human individuals in terms of an embedded frontal lobe, aggregated neural network signals or the laws of chemistry respectively. Now, there are certainly cases where the extended, collective or artificial epistemic agent and its abilities can be systematically mapped onto an aggregate of its embedded individual actions (e.g. Lenny), its member contributions (e.g. Like-Minded Summative Class) or its algorithmic procedures (e.g. APLI), thus making the additional epistemic agent an explanatorily superfluous postulate, but the objection that extended, collective or artificial agency is automatically superfluous mistakes supervenience for explanatory reducibility by failing to acknowledge that macro-systematicities have *explanatory power* (e.g. in shorthands, where the macro-systematicites are convenient, even if they can systematically be redescribed as micro-systematicities) and can be *explanatorily unique* (e.g. in longhands, where the lack of systematic mapping-relation due to assembly bonus effects would force us to give up the macro-systematicities in favour of long winded redescriptions that apply only to particular cases at particular times). If the epistemic stance targeting the coupled system, group or artificial system lacks a straightforward mapping onto the epistemic stance of its components (e.g. Coqto, Expert Planet, Deep Blue), then what makes the emerging epistemic agent a uniquely powerful explanatory postulate would get lost in a reducibility-story.

The *regress problem* was the worry or criticism that the epistemic agent postulate is *superfluous* because its abilities and epistemic properties *regress* to causes *outside of the agent*. But the abilities and epistemic agency of artificial systems can be as difficult to account for in terms of the abilities or epistemic agency of its programmers as is the case with human students and their teachers. Now, there are certainly cases where this regress can be claimed as a defeater of the epistemic agency (e.g. puppets, letters) or a superior reading of the situation (e.g. AAS) because nothing explanatory is lost

in the regress. After all, if we can straightforwardly map the acts or epistemic properties of the (artificial) epistemic agent to its origin(s), then that epistemic agent postulate provides nothing explanatorily new or distinct over the regression to that of its origin(s). But the claim that any postulate of artificial epistemic agency must regress takes the legitimate worry of a superfluous epistemic stance, and unduly extends it to any case where there is a causal origin (e.g. teaching or programming), thereby failing to acknowledge that the explanation invoking the epistemic agent can be *explanatorily self-sufficient* (e.g. even though each of AAS's answers regress to Ohce, the program functions independently of her) or *explanatorily distinct* (e.g. regressing Deep Blue's game could only be done by reducing each individual decision to the particular algorithmic procedures invoked, and regressing those to the programmers who wrote them). If the epistemic stance targeting the artificial system lacks a straightforward mapping onto the epistemic stance of its author(s) (e.g. Deep Blue, Adam), then what makes the epistemic agent a uniquely powerful explanatory postulate would get lost in a regress-story. This means there is no benefit, and even a disadvantage, to changing the explanation from the program to its programmer. So it is not fair to say that the agency or abilities revealed by each program always regress to those of its programmers in the same way that the agency revealed by letters or puppets regress. Furthermore, the complexity of the mapping relation is a difference in degree, and not kind. It is true that this entails there is no ultimacy of authorship, but we readily acknowledge that human achievements are not borne ex nihilo, so why must we hold artificial agents to a higher standard?

Lastly, the *rigidity problem* was the worry or criticism that artificial systems are inherently limited in their abilities due to being a *rigid* formal system. If understanding a certain scientific topic is composed of *informal* abilities (meaning they cannot be exhausted by a finite formalised rule or set of rules), then any entity that relies on a formalised set of abilities, such as an artificial system, will lack in quality (sensitivity and scope in particular). This seems plausible, given that many of even the best examples of candidates for artificial understanding find their abilities in brute-force calculation techniques, and seem to display (mostly) less advanced understanding than we would find in human experts. Nevertheless, the rigidity problem mistakenly assumes that the level of computation must align with the level of informal abilities, thereby failing to acknowledge that *the computational level* may fall well below that level. As we saw in discussing reducibility and emergence, assembly bonuses may emerge from limited low level systematicities. Furthermore, these assembly bonuses need not themselves be capturable by formalised rules. And since epistemic thought-processes emerge from the architecture of the brain, and since they furthermore appear to defy formalization to such an extent, it may be that we will need to focus on formalising the appropriate sub-cognitive processes if we want to create

artificial scientists. There is hope for emergent computation in the form of a decentralised system composed of dynamically cooperating sub-processes forming global learning and flexible abilities. Given this, it is my contention that we have no reason to suspect that the possible advancements of automating scientists are soon to be exhausted.

As such, I have presented a mark of understanding and a corresponding mark of epistemic subjecthood that allows us to successfully talk about the features, quality, evaluations, problems and candidate subjects in a philosophically coherent and interdisciplinarily justified way. I hope to have herewith shown that my account is a valuable contribution in characterising understanding and the understanding subject.

DISSERTATION

Reference List

Aberdein, A. (2006). Managing Informal Mathematical Knowledge: Techniques from informal logic. In Borwein, J.M. & Farmer, W.M. (Eds.), *Mathematical Knowledge Management* (pp. 208–221). Berlin: Springer.

Aberdein, A. (2007). The Informal Logic of Mathematical Proof. In B. Van Kerkhove B. & J. P. Van Bendegem (Eds.), *Perspectives on Mathematical Practices* (pp. 135–151). Dordrecht: Springer.

Adams, F., & Aizawa, K. (2001). The Bounds of Cognition. *Philosophical psychology,* 14(1), 43-64.

Adams, F., & Aizawa, K. (2010). Defending the Bounds of Cognition. In R. Menary (Ed.), *The Extended Mind* (pp. 67–80). Cambridge, MA: MIT Press.

Allo, P., Van Bendegem, J. P., & Van Kerkhove, B. (2013). Mathematical Arguments and Distributed Knowledge. In A. Aberdein & I.J. Dove (Eds.), *The Argument of Mathematics* (pp. 339-360). New York: Springer.

Avigad, J. (2008). Understanding Proof. In P. Mancosu (Ed.), *The Philosophy of Mathematical Practice* (pp. 317-353). New York: Oxford University Press.

Avigad, J. (2010). Understanding, Formal Verification, and the Philosophy of Mathematics. *Journal of the Indian Council of Philosophical Research,* 27, 161-197.

Azzouni, J. (2004). The Derivation-indicator View of Mathematical Practice. *Philosophia Mathematica*, 12(2), 81-106.

Baber, C., Smith, P., Cross, J., Hunter, J. E., & McMaster, R. (2006). Crime Scene Investigation as Distributed Cognition. *Pragmatics & Cognition*, 14(2), 357-385.

Baker, A. (2015). Non-deductive Methods in Mathematics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from http://plato.stanford.edu/archives/fall2015/entries/mathematics-nondeductive

Barber, A. (Ed.) (2003). *Epistemology of Language*. Oxford, UK: Oxford University Press.

Barmby, P., Harries, T., Higgins, S., and Suggate, J. (2007). How Can We Assess Mathematical Understanding? In J. H. Woo, H. C. Lew, K. S. Park, and D. Y. Seo (Eds.), *Proceedings of the 31st Conference of the International Group for the Psychology of Mathematics Education Vol. 2* (pp. 41-48). Seoul, South Korea: PME.

Barthes, R. (2001). The Death of the Author. *Contributions in Philosophy*, 83, 3-8.

Baumberger, C. (2011). Understanding and its Relation to Knowledge. In C. Jäger and W. Löffler (Eds.), *Epistemology: Contexts, Values, Disagreement. Papers of the 34th International Wittgenstein Symposium* (pp. 16-18).

Baumberger, C., Beisbart, C., and Brun, G. (2016). What is Understanding? An overview of recent debates in epistemology and philosophy of science. In S.R. Grimm, C. Baumberger, and S. Ammon (Eds.), *Explaining Understanding: New perspectives from epistemology and philosophy of science* (pp. 1–34). New York: Routledge.

Bearman, S., Korobov, N., & Thorne, A. (2009). The fabric of Internalized Sexism. *Journal of Integrated Social Sciences*, 1(1), 10-47.

Benzmüller, C., Jamnik, M., Kerber, M., & Sorge, V. (1999). Agent Based Mathematical Reasoning. *Electronic Notes in Theoretical Computer Science*, 3(23), 340-351.

Benzmüller, C., Kerber, M., Jamnik, M., & Sorge, V. (2001). Experiments with an Agent-oriented Reasoning System. In F. Baader, G. Brewka, & T. Eiter (Eds.), *KI 2001: Advances in Artificial Intelligence*. (pp. 409-424). Berlin: Springer.

Bickle, J. (2019). Multiple Realizability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/spr2019/entries/multiple-realizability/

Bird, A. (1998). Dispositions and Antidotes. *The Philosophical Quarterly,* 48(191), 227-234.

Bird, A. (2014). When is There a Group that Knows? Distributed cognition, scientific knowledge, and the social epistemic subject. In J. Lackey (Ed.), *Essays in Collective Epistemology* (pp. 42-63). New York: Oxford University Press.

Block, N. (1981). Psychologism and Behaviorism. *The Philosophical Review,* 90(1), 5-43.

Block, N. J., & Fodor, J. A. (1972). What Psychological States Are Not. *The Philosophical Review*, 81(2), 159-181.

Block, N., (1978). Troubles With Functionalism. In. C.W. Savage (Ed.), *Perception and Cognition: Issues in the Foundations of Psychology. Minnesota Studies in the Philosophy of Science, Vol. 9* (pp. 261-325). Minneapolis: University of Minnesota Press. Retrieved from https://web.archive.org/web/20110927150409/http://w3.uniroma1.it/cordeschi/Articoli/block.htm

Borowiec, S. (2016). AlphaGo Seals 4-1 Victory over Go Grandmaster Lee Sedol. *The Guardian.* Retrieved from https://www.theguardian.com/technology/2016/mar/15/googles-alphago-seals-4-1-victory-over-grandmaster-lee-sedol#

Boyd, K. (2019). Group Understanding. *Synthese.* Retrieved from https://philpapers.org/archive/BOYGU.pdf

Bratman, M. E. (2013). *Shared agency: A planning theory of acting together.* Oxford: Oxford University Press.

Breyer, D., & Greco, J. (2008). Cognitive Integration and the Ownership of Belief: Response to Bernecker. *Philosophy and Phenomenological Research*, 76(1), 173-184.

Bringsjord, S. (2001). Creativity, the Turing Test, and the (Better) Lovelace Test. *Minds and Machines*, 11, 3-27. Retrieved from http://kryten.mm.rpi.edu/lovelace.pdf

Brogaard, B. (2005). *I Know. Therefore, I Understand* (Unpublished typescript).

Brown, H. I. (1988). *Rationality.* London: Routledge.

Burge, T. (1979). Individualism and the Mental. *Midwest studies in philosophy,* 4(1), 73-121.

Burge, T. (1998). Computer Proof, Apriori Knowledge, and Other Minds. *Noûs*, 32, 1–37.

Carroll, L. (1895). What the Tortoise Said to Achilles. *Mind*, 4(14), 278-280.

Carter, J. A. & Gordon, E. C. (forthcoming). On Pritchard, Objectual Understanding and the Value Problem. *American Philosophical Quarterly*. Retrieved from

https://www.academia.edu/3024604/On_Pritchard_Objectual_Understanding_and_the_Value_Problem

Carter, J. A., Clark, A., Kallestrup, J., Palermos, S. O., & Pritchard, D. (Eds.). (2018). *Extended epistemology.* Oxford, UK: Oxford University Press.

Castlevecchi, D. (2015). Physics Paper Sets Record with More Than 5,000 Authors. *Nature.* Retrieved from https://www.nature.com/news/physics-paper-sets-record-with-more-than-5-000-authors-1.17567

Cellucci, C. (2000). The Growth of Mathematical Knowledge: An open world view. In E. Grosholz & H. Breger (Eds.), *The Growth of Mathematical Knowledge* (pp. 153-176). Dordrecht: Kluwer.

CERN. (2012). CERN Experiments Observe Particle Consistent with Long-sought Higgs boson (Press Release). CERN. Retrieved from https://home.cern/news/press-release/cern/cern-experiments-observe-particle-consistent-long-sought-higgs-boson

Chalmers, D. J. (1990). Computing the Thinkable. *Behavioral and Brain Sciences*, 13(04), 658-659.

Chang, H. (2012). *Is water H2O? Evidence, realism and pluralism.* Dordrecht: Springer

Choi, S. (2008). Dispositional Properties and Counterfactual Conditionals. *Mind,* 117(468), 795-841.

Choi, S. (2011). What is a Dispositional Masker? *Mind,* 120(480), 1159-1171.

Choi, S. & Fara, M. (2018). Dispositions. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/fall2018/entries/dispositions/

Chomsky, N. (1997). Language and Problems of Knowledge. *Teorema,* 16(2), 5-33.

Churchland, P. M. (1992). *A nNurocomputational Perspective: The nature of mind and the structure of science.* Cambridge, MA: MIT press.

Clark, A. (2008). *Supersizing the Mind: Embodiment, action, and cognitive extension.* New York: Oxford. University Press.

Clark, A. (2010). Memento's Revenge: The extended mind, extended. *The extended mind*, 43-66.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1), 7-19.

Coates, D. J., & McKenna, M. (2015). Compatibilism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from http://plato.stanford.edu/entries/compatibilism/

Cole, D. (2014). The Chinese Room Argument. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/spr2020/entries/chinese-room/

Collins, B. E., & Guetzkow, H. S. (1964). *A Social Psychology of Group Processes for Decision-making.* New York: Wiley. Retrieved from https://archive.org/details/socialpsychology0000coll

Colton, S., Bundy, A., & Walsh, T. (1999). HR: Automatic concept formation in pure mathematics. In T. Dean (Ed.), *Proceedings of the 16th International Joint Conference on Artificial Intelligence* (pp. 786–791). San Francisco: Morgan Kaufmann Publishers.

Colton, S., Bundy, A., & Walsh, T. (2000). Agent based cooperative theory formation in pure mathematics. In G. Wiggins (Ed.), Proceedings of AISB 2000 Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science, (pp. 11–18). Birmingham, UK

Davis, P. J., & Hersh, R. (1998). The Ideal Mathematician. In T. Tymoczko (Ed.), *New Directions in the Philosophy of Mathematics: revised and expanded* (pp. 177-184). New Jersey: Princeton University Press.

Dawson, J. W. (2006). Why Do Mathematicians Re-prove Rheorems? *Philosophia Mathematica*, 14(3) 269-286.

de Regt, H. W. (2009). The Epistemic Value of Understanding. *Philosophy of Science*, 76(5), 585-597.

de Regt, H. W. (2004). Discussion Note: Making sense of understanding. *Philosophy of Science*, 71(1), 98-109.

de Regt, H. W. (2017). Understanding Scientific Understanding. New York: Oxford University Press.

de Regt, H. W., & Dieks, D. (2005). A Contextual Approach to Scientific Understanding. *Synthese*, 144(1), 137-170.

de Regt, H. W., & Gijsbers, V. (2016). How False Theories Can Yield Genuine Understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding* (pp. 50–75). London: Routledge.

de Ridder, G. J. (2018). Representations and Robustly Collective Attitudes. In J. A. Carter, A. Clark, J. Kallestrup, O. Palermos, & D. Pritchard (Eds.), *Socially Extended Epistemology* (pp. 36-58). Oxford: Oxford University Press.

de Rosnay, J. (1979). *The Macroscope: A new world scientific system*. New York: HarperCollins Publishers. Retrieved from http://cleamc11.vub.ac.be/macroscope/

Delarivière, S. (2015). *Artificial Free Will: In which an account of free will is argued for that is compatible with determinism, correspondent with consciousness and constructible for artificial intelligence* (MA Thesis). Vrije Universiteit Brussel.

Delarivière, S. (2016). Artificial Free Will: The Responsibility Strategy and Artificial Agents. *Apeiron Student Journal of Philosophy (Portugal),* 7(1), 175-203.

Delarivière, S. (2020). Collective Understanding: A conceptual defense for when groups should be regarded as epistemic agents with understanding. *AVANT*, 12(2).

Delarivière, S., & Van Kerkhove, B. (2017). The "Artificial Mathematician" Objection: Exploring the (Im)possibility of Automating Mathematical Understanding. In B. Sriraman (Ed.), *Humanizing Mathematics and its Philosophy* (pp. 173-198). Basel: Birkhäuser.

Delarivière, S., Frans, J., & Van Kerkhove, B. (2017). Mathematical Explanation: A contextual approach. *Journal of Indian Council of Philosophical Research*, 34(2), 309-329.

Dennett, D. C. (1978a). A cure for the common code. In D. C. Dennett (Ed.), *Brainstorms: Philosophical essays on mind and psychology* (pp. 99-118). Cambridge, MA: MIT Press.

Dennett, D. C. (1978b). Where am I? In D. C. Dennett (Ed.), *Brainstorms: Philosophical essays on mind and psychology* (pp. 333–346). Cambridge, MA: MIT Press.

Dennett, D. C. (1978c). Toward a Cognitive Theory of Consciousness. In D. C. Dennett (Ed.), *Brainstorms: Philosophical essays on mind and psychology* (pp. 163-188). Cambridge, MA: MIT Press.

Dennett, D. C. (1990). *The Intentional Stance.* Cambridge MA: MIT Press.

Dennett, D. C. (1993). *Consciousness Explained.* London: Penguin Books.

Dennett, D. C. (1997). When Hal Kills, Who's to Blame? Computer ethics. In D. G. Stork (Ed.), *Hal's Legacy: 2001's Computer as Dream and Reality* (pp. 351-365). Cambridge, MA: MIT Press.

Dennett, D. C. (1998). The logical geography of computational approaches: A view from the East Pole. In D. C. Dennett. (Ed.), *Brainchildren: Essays on designing minds* (pp. 215–234). London: Penguin Books. (Original work published 1986).

Dennett, D. C. (1998). The Myth of Double Transduction. In S. Hameroff, A. W. Kaszniak, & A. C. Scott (Eds.), *The International Consciousness Conference: Toward a science of consciousness II, the second Tucson discussions and debates* (pp. 97–107). Cambridge, MA: MIT Press.

Dennett, D. C. (Ed.). (1998). *Brainchildren: Essays on designing minds.* London: Penguin Books.

Dennett, D. C. (2004). *Freedom Evolves.* London: Penguin Books.

Dennett, D. C. (2009). Intentional Systems Theory. In Beckermann, A., McLaughlin, B. P., Walter, S. (Eds.), *The Oxford Handbook of Philosophy of Mind* (pp. 339–350). Oxford, UK: Oxford University Press.

Dennett, D. C. (2013). *Intuition Pumps and Other Tools for Thinking.* London: Penguin Books.

Dennett, D. C., & Hofstadter, D. R. (Eds.). (1985). *The Mind's I: Fantasies and reflections on self and soul.* Middlesex, England: Penguin Books.

Dennett, D. C., & Taylor, C. (2002). Who's Afraid of Determinism? Rethinking Causes and Possibilities. In R. Kane (Ed.), *The Oxford Handbook of Free Will* (pp. 257-277). New York: Oxford University Press.

Detlefsen, M. & Luker, M. (1980). The Four-Color Theorem and Mathematical Proof. *The Journal of Philosophy,* 77(12), 803-820.

Dewey, J. (1933). *How We Think: A restatement of the relation of reflective thinking to the educative process.* Boston: DC Heath.

Dutant, J. (2015). The Legend of the Justified True Belief Analysis. *Philosophical Perspectives*, 29(1), 95-145.

Ekstrom, L. W. (2005). Alienation, Autonomy, and the Self. *Midwest studies in philosophy,* 29, 45-67.

Elgin, C. Z. (2007). Understanding and the Facts. *Philosophical Studies,* 132(1), 33-42.

Elgin, C. Z. (2009). Is Understanding Factive? In A. Haddock, A. Millar & D. Pritchard (Eds.), *Epistemic Value* (pp. 322–30). Oxford: Oxford University Press.

Elgin, C. Z. (2004). True Enough. *Philosophical issues,* 14, 113-131.

Epstein, B. (2017). What are Social Groups? Their metaphysics and how to classify them. *Synthese.* Retrieved from https://www.readcube.com/articles/10.1007/s11229-017-1387-y

Epstein, B. (2018). Social Ontology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/sum2018/entries/social-ontology/

Fantl, J. (2017). Knowledge How. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/fall2017/entries/knowledge-how/

Fara, M. (2008). Masked abilities and compatibilism. *Mind*, 117(468), 843-865.

Fischer, J. & Ravizza, M. (1998). *Responsibility and Control. A theory of moral responsibility.* New York: Cambridge University Press.

Fischer, J. M. (2002). Frankfurt-type Examples and Semi-Compatibilism. In R. Kane (Ed.), *The Oxford Handbook of Free Will* (pp. 281-308). New York: Oxford University Press.

Ford, K. M., Glymour, C., & Hayes, P. J. (Eds.). (1995). *Android Epistemology.* Cambridge, MA: MIT Press.

Ford, K. M., Glymour, C., & Hayes, P. J. (Eds.). (2006). Thinking about android epistemology. Cambridge, MA: MIT Press.

Forrest, S. (1991). Emergent Computation: self-organizing, collective, and cooperative phenomena in natural and artificial computing networks. In S. Forrest (Ed.), *Emergent computation* (p. 1-11). Cambridge, MA: MIT Press.

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy*, 66(23), 829-839

Frans, J. (2020). Unificatory Understanding and Explanatory Proofs. *Foundations of Science,* 1-23.

Frans, J., & Weber, E. (2014). Mechanistic Explanation and Explanatory Proofs in Mathematics. *Philosophia Mathematica,* 22(2), 231-248.

Friedman, M. (1974). Explanation and Scientific Understanding. *The Journal of Philosophy*, 71(1), 5-19.

Geirsson, H. (2004). Contra Collective Epistemic Agency. *Southwest Philosophy Review,* 20(2), 163-166. Retrieved from https://pdfs.semanticscholar.org/b833/c3d00c4640527dc7cf8de1844036ef4d2811.pdf

Geist, C., Löwe, B., & Van Kerkhove, B. (2010). Peer Review and Knowledge by Testimony in Mathematics. In B. Löwe and T. Müller (Eds.), *Philosophy of Mathematics: Sociological Aspects and Mathematical Practice* (pp. 155-178). London: College Publications.

Gettier, E. L. (1963). Is Justified True Belief Knowledge? *Analysis,* 23(6), 121-123.

Giere, R. N. (2002a). Models as Parts of Distributed Cognitive Systems. In In L. Magnani & N. J. Nersessian (Eds.), *Model-based reasoning* (pp. 227-241). Boston: Springer.

Giere, R. N. (2002b). Scientific Cognition as Distributed Cognition. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The Cognitive Basis of Science* (p. 285–299). Cambridge: Cambridge University Press.

Giere, R. N. (2002c). Discussion Note: Distributed cognition in epistemic cultures. *Philosophy of Science,* 69(4), 637-644. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.367.136&rep=rep1&type=pdf

Gilbert, M. (2000). Collective Belief and Scientific Change. In M. Gilbert (Eds.), *Sociality and Responsibility: New essays on plural subject theory* (pp. 37–49). Lanham, MD: Rowman & Littlefield

Gilbert, M. (2004). Collective Epistemology. *Episteme*, 1(2), 95-107.

Gilbert, M. (2013). *Joint Commitment: How We Make the Social World.* New York: Oxford University Press.

Glick, E. (2012). Abilities and Know-How Attributions. In J. Brown & M. Gerken (Eds.*), Knowledge Ascriptions* (pp. 120–139). Oxford: Oxford University Press.

Godino, J. D. (1996). Mathematical Concepts, Their Meanings, and Understanding. In L. Puig and A. Gutierrez (Eds.), *Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education Vol. 2* (pp. 417-424). University of Valencia.

Goldin, G. A. (1998). Representational Systems, Learning, and Problem Solving in Mathematics. *The Journal of Mathematical Behavior*, 17(2), 137-165.

Goldman, A. & Beddor, B. (2016). Reliabilist Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/win2016/entries/reliabilism/

Goldman, Al. & Blanchard, T. (2018). Social Epistemology. In E. N. Zalte (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2018/entries/epistemology-social

Gopnik, A. (1998). Explanation as Orgasm. *Minds and Machines*, 8(1), 101-118.

Gordon (n.d.). Understanding in Epistemology. In *Internet Encyclopedia of Philosophy*. Retrieved from https://www.iep.utm.edu/understa/

Gordon, E. C. (2012). Is There Propositional Understanding? *Logos & Episteme*, 3(2), 181-192.

Greco, J. (2000). *Putting Skeptics in Their Place: The nature of skeptical arguments and their role in philosophical inquiry*. Cambridge: Cambridge University Press.

Greco, J. (2007). The Nature of Ability and the Purpose of Knowledge. *Philosophical Issues*, 17(1), 57-69.

Greco, J. (2008). What's Wrong With Contextualism? *The Philosophical Quarterly*, 58(232), 416-436.

Griffin, J. (1986). *Well-Being: Its meaning, measurement and moral importance*. Oxford, UK: Clarendon Press.

Grimm, S. R. (2006). Is Understanding a Species of Knowledge? *The British Journal for the Philosophy of Science*, 57(3), 515-535.

Grimm, S. R. (2011). Understanding. In D. Pritchard & S. Bernecker (Eds.), *The Routledge Companion to Epistemology* (pp. 84–94). New York: Routledge.

Grimm, S. R. (2012). The Value of Understanding. Philosophy *Compass*, 7(2), 103-117. Retrieved from https://faculty.fordham.edu/sgrimm/Site/Papers_and_Presentations_files/value%20of%20understanding-philosophy%20compass-8-14-11.pdf

Grimm, S. R. (2014). Understanding as Knowledge of Causes. In A. Fairweather (Ed.), *Virtue epistemology naturalized* (pp. 329-345). Dordrecht: Springer.

Grimm, S. R. (2016). Understanding and Transparency. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining Understanding: New perspectives from epistemology and philosophy of science* (pp. 251–271). London: Routledge

Hales, T. (2008). Formal Proof. *Notices of the AMS*, 55(11), 1370-1380

Hanks, P., & Hanks, P. W. (2015). *Propositional content*. Oxford: Oxford University Press.

Hansen, H. (2019). Fallacies. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2019/entries/fallacies/

Harris, S. (2012). *Free Will*. New York: Free Press.

Harrison, J. (2008). Formal Proof - Theory and Practice. *Notices of the AMS*, 55(11), 1395-1460.

Haugeland, J. (1979). Understanding Natural Language. *The Journal of Philosophy*, 76(11), 619-632.

Hempel, C. (1965). *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science*. New York: Free Press.

Hernandez, E., Sanchez-Anguix, V., Julian, V., Palanca, J., & Duque, N. (2016). Rainfall Prediction: A Deep Learning Approach. In F. Martínez-Álvarez, A. Troncoso, H. Quintián, & E. Corchado (Eds.), *International Conference on Hybrid Artificial Intelligence Systems* (pp. 151-162). Switzerland: Springer Verlag.

Hersh, R. (1997). *What is Mathematics, Really?* Oxford: Oxford University Press.

Heylighen, F. (2014). *Cognitive Systems: A cybernetic perspective on the new science of the mind* (Lecture Notes). Retrieved from http://pespmc1.vub.ac.be/papers/cognitivesystems.pdf

Hiebert, J., & Carpenter, T. P. (1992). Learning and Teaching with Understanding. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning: A project of the National Council of Teachers of Mathematics* (pp. 65–97). New York: Macmillan.

Hills, A. (2009). Moral Testimony and Moral Epistemology. *Ethics,* 120(1), 94-127.

Hills, A. (2015). Understanding Why. *Noûs*, 50, 661–688. Retrieved from https://www.academia.edu/9860382/Understanding_why

Hofstadter, D. R. (1982). Artificial Intelligence: Subcognition as computation. *Technical Report 132, Computer Science Department. Indiana University.* Retrieved from https://legacy.cs.indiana.edu/ftp/techreports/TR132.pdf

Hofstadter, D. R. (1999). *Gödel, Escher, Bach: An eternal golden braid.* New York: Basic Books.

Huebner, B. (2013). *Macrocognition: A theory of distributed minds and collective intentionality.* New York: Oxford University Press.

Hutchins, E. (1995). *Cognition in the Wild.* Cambridge: MIT press.

Internalized Oppression. (2019). In *Wikipedia, The Free Encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Internalized_oppression&oldid=920199220

Iten, R., Metger, T., Wilming, H., Del Rio, L., & Renner, R. (2020). Discovering Physical Concepts with Neural Networks. *Physical Review Letters,* 124(1), 010508. Retrieved from https://arxiv.org/pdf/1807.10300.pdf

Japyassú, H. F., & Laland, K. N. (2017). Extended Spider Cognition. *Animal Cognition*, 20(3), 375-395.

Johnston, M. (1992). How to Speak of the Colors. *Philosophical Studies*, 68(3), 221-263.

Jubien, M. (2001). Propositions and the Objects of Thought. *Philosophical Studies*, 104(1), 47-62.

Kane, R. (2002). Introduction: The contours of contemporary free will debates. In R. Kane (Ed.), *The Oxford Handbook of Free Will* (pp. 3-41). New York: Oxford University Press.

Kelp, C. (2015). Understanding Phenomena. *Synthese,* 192(12), 3799-3816. Retrieved from http://eprints.gla.ac.uk/140962/7/140962.pdf

Khalifa, K. (2012). Inaugurating Understanding or Repackaging Explanation? *Philosophy of Science,* 79(1), 15-37.

Khalifa, K. (2013). Understanding, Grasping and Luck. *Episteme*, 10(1), 1-17.

Kim, J. (1992). Multiple Realization and the Metaphysics of Reduction. *Philosophy and Phenomenological Research*, 52(1), 1-26.

King, R. D., Rowland, J., Aubrey, W., Liakata, M., Markham, M., Soldatova, L. N., Whelan, K. E., Clare, A., Young, M., Sparkes, A., Oliver, S. G., Pir, P. (2009). The Robot Scientist Adam. *Computer*, 42(8), 46-54.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G., Bryant, C. H., Muggleton, S. H., Kell, D.B., & Oliver, S. G. (2004). Functional Genomic Hypothesis Generation and Experimentation by a Robot Scientist. *Nature,* 427(6971), 247-252.

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher and W. C. Salmon (Eds.), *Scientific Explanation* (pp. 410–505). Minneapolis: University of Minnesota Press

Knorr Cetina, K. (1999). *Epistemic Cultures: The cultures of knowledge societies*. Cambridge, MA: Harvard University Press.

Kuorikoski, J. (2011). Simulation and the Sense of Understanding. In P. Humphreys, & C. Imbert (Eds.), *Models, Simulations, and Representations* (pp. 168-187). New York: Routledge. Retrieved from http://philsci-archive.pitt.edu/4480/1/simulationandsou.pdf

Kuorikoski, J., & Ylikoski, P. (2015). External Representations and Scientific Understanding. *Synthese,* 192(12), 3817-3837.

Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: Cambridge University Press.

Lakatos, I. (1976). *Proofs and Refutations: the logic of mathematical discovery*. Cambridge: Cambridge University Press.

Lakoff, G., & Núñez, R. E. (2000). *Where Mathematics Comes From: How the embodied mind brings mathematics into being*. New York: Basic Books.

Lange, M. (2014). Aspects of mathematical explanation: Symmetry, unity, and salience. Philosophical Review, 123(4), 485–531.

Larson, C. E. (2005). A Survey of Research in Automated Mathematical Conjecture-Making. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 69, 297–318.

Laughlin, P. R., Hatch, E. C., Silver, S., Boh, L. (2006). Groups Perform Better Than the Best Individuals on Letters-to-Numbers Problems: Effects of group size. *Journal of Personality and Social Psychology*. 90 (4): 644–651.

Leroy, X. (2014). *Formal Verification of a Static Analyzer: Abstract interpretation in type theory* (Powerpoint slides). Retrieved from https://xavierleroy.org/talks/TYPES-2014.pdf

Lewis, D. (1973). Counterfactuals and Comparative Possibility. *Journal of Philosophical Logic* 2(4), 418–446.

Lewis, D. (1997). Finkish Dispositions. *The Philosophical Quarterly*, 47(187), 143-158.

Lipton, P. (2009). Understanding without Explanation. In H. W. de Regt, S. Leonelli & K. Eigner (Eds.), *Scientific Understanding: Philosophical perspectives*. University of Pittsburgh Press. p. 43-63.

List, C., & Pettit, P. (2006). Group Agency and Supervenience. *The Southern Journal of Philosophy,* 44(S1), 85-105.

List, C., & Pettit, P. (2011). *Group Agency: The possibility, design, and status of corporate agents*. Oxford, UK: Oxford University Press.

Luper, S. (2020). Epistemic Closure. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/sum2020/entries/closure-epistemic/

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2014). Traffic Flow Prediction with Big Data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2), 865-873.

Macbeth, D. (2012). Proof and Understanding in Mathematical Practice. *Philosophia Scientiæ. Travaux d'histoire et de philosophie des sciences,* 16(1), 29-54.

MacKenzie, D. (1999). Slaying the Kraken: The Sociohistory of a Mathematical Proof. *Social Studies of Science,* 29(1), 7-60.

MacKenzie, D. (2004). *Mechanizing Proof: Computing, risk, and trust.* Cambridge: MIT Press.

Mancosu, P. (2008). *The Philosophy of Mathematical Practice.* Oxford: Oxford University Press.

Manley, D., & Wasserman, R. (2007). A Gradable Approach to Dispositions. *The Philosophical Quarterly*, 57(226), 68-75.

Marie Curie. (2020). In *Wikiquote*. Retrieved from https://en.wikiquote.org/w/index.php?title=Marie_Curie&oldid=2741000.

Mathiesen, K. (2006). The Epistemic Features of Group Belief. *Episteme,* 2(3), 161-175.

McCune, W. (1990). Otter 2.0 (theorem-prover). In M. E. Stickel (Ed.), *Proceedings of the 10th International Conference on Automated Deduction* (pp. 663–664). Berlin: Springer-Verlag.

Menzies, P. (2014). Counterfactual Theories of Causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from http://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/

Milkowski, M. (2013). *Explaining the Computational Mind.* Cambridge: MIT Press.

Miłkowski, M. (2017). Situatedness and Embodiment of Computational Systems. *Entropy*, 19(4), 162.

Mitchell, M., & Hofstadter, D. R. (1990). The Emergence of Understanding in a Computer Model of Concepts and Analogy-Making. *Physica D,* 42(1-3), 322-334.

Moore, G. E. (1912). Free Will. In G.E. Moore (Ed.), *Ethics* (pp. 84-95). Oxford: Oxford University Press.

Morris, K. (2011). A Defense of Lucky Understanding. *The British Journal for the Philosophy of Science,* 63(2), 357-371.

Newman, M. (2012). An Inferential Model of Scientific Understanding. *International Studies in the Philosophy of Science,* 26(1), 1-26.

O'Conor, T. & Wong, H. Y. (2015). Emergent Properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from http://plato.stanford.edu/entries/properties-emergent/

Oppy, G., & Dowe, D. (2011). The Turing Test. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from http://plato.stanford.edu/archives/spr2011/entries/turing-test/

Palermos, O. & Pritchard, D. (2016). The Distribution of Epistemic Agency. In P. Reider (Ed.), *Social Epistemology and Epistemic Agency: De-centralizing epistemic agency* (pp. 109-126). London: Rowman & Littlefield.

Palermos, S. O. (2016). The Dynamics of Group Cognition. *Minds and Machines*, 26(4), 409-440.

Palermos, S. O., & Pritchard, D. (2013). Extended Knowledge and Social Epistemology. *Social Epistemology Review and Reply Collective*, 2(8), 105-120.

Pavitt, C. (2003). Colloquy: Do interacting groups perform better than aggregates of individuals? Why we have to be reductionists about group memory. *Human Communication Research*, 29(4), 592-599.

Peacocke, C. (1983). *Sense and Content.* Oxford: Oxford University Press.

Pease, A. (2007). *A Computational Model of Lakatos-style Reasoning* (Doctoral dissertation). University of Edinburgh.

Pease, A., Crook, P., Smaill, A., Colton, S., & Guhe, M. (2009). Towards a Computational Model of Embodied Mathematical Language. In Proceedings of the Second Symposium on Computing and Philosophy (pp. 35-37). Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Pease, A., Smaill, A., Colton, S., Ireland, A., Teresa Llan, M., Ramezani, R., Grov, G., & Gube, M. (2010). Applying Lakatos-style Reasoning to AI Problems. In J. Vallverdú (Ed.), *Thinking Machines and the Philosophy of Computer Science: Concepts and Principles* (pp. 149-174). Hershey: IGI Global.

Penrose, R. (1999). *The Emperor's New Mind.* Oxford: Oxford University Press.

Pereboom, D. (2006). *Living Without Free Will.* Cambridge: Cambridge University Press.

Pettit, P. (2003). *Groups With Minds of Their Own*. In F. Schmitt (Ed.), *Socializing Metaphysics* (p. 167-193). New York: Rowan and Littlefield.

Pettit, P. (2007). Rationality, Reasoning and Group Agency. *Dialectica*, 61(4), 495-519.

Pettit, P. (2014). How to Tell if a Group is an Agent. In J. Lackey (Ed.), *Essays in Collective Epistemology* (pp. 97-121). Oxford: Oxford University Press.

Poston, T. (2009). Know How to be Gettiered? *Philosophy and Phenomenological Research*, 79(3), 743-747.

Praino, A. P., Treinish, L. A., Christidis, Z. D., & Samuelsen, A. (2003). Case Studies of an Operational Mesoscale Modelling System in the Northeast United States. In *Proceedings of the Nineteenth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography and Hydrology*.

Pritchard, D. (2005). *Epistemic Luck*. New York: Oxford University Press.

Pritchard, D. (2008). Sensitivity, Safety, and Anti-Luck Epistemology. In J. Greco (Ed.), *The Oxford Handbook of Scepticism* (pp. 437-455). Oxford: Oxford University Press.

Pritchard, D. (2009). Knowledge, Understanding and Epistemic Value. *Royal Institute of Philosophy Supplement,* 64, 19-43.

Pritchard, D. (2010). Cognitive Ability and the Extended Cognition Thesis. *Synthese*, 175(1), 133-151.

Pritchard, D. (2014). Knowledge and Understanding. In A. Fairweather (Ed.), *Virtue epistemology naturalized* (pp. 315-327). Cham: Springer.

Putnam, H. (1975). The Meaning of "Meaning". In K. Gunderson (Ed.), *Language, Mind and Knowledge: Minnesota Studies in the Philosophy of Science.* Minneapolis: University of Minnesota Press.

Quine, W. V. O. (1990). *Pursuit of Truth.* Cambridge, MA: Harvard University Press.

Rav, Y. (1999). Why Do We Prove Theorems? *Philosophia Mathematica*, 7(3), 5-41.

Reber, A. S. (1989). Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology:* General, 118(3), 219.

Riggs, W. D. (2003). Understanding "Virtue" and the Virtue of Understanding, In M. R. DePaul and L. Zagzebski (Eds.), *Intellectual Virtue: Perspectives from ethics and epistemology* (pp. 203-226). Oxford: Clarendon Press.

Rupert, R. D. (2004). Challenges to the Hypothesis of Extended Cognition. *The Journal of Philosophy,* 101(8), 389-428.

Rupert, R. D. (2005). Minding One's Cognitive Systems: When does a group of minds constitute a single cognitive unit? *Episteme,* 1(3), 177-188.

Rupert, R. D. (2009). *Cognitive Systems and the Extended Mind.* Oxford University Press.

Rupert, R. D. (2011). Empirical Arguments for Group Minds: A critical appraisal. *Philosophy Compass*, 6(9), 630-639.

Rupert, R. D. (2013). Distributed Cognition and Extended-Mind Theory. In B. Kaldis (Ed.), *Encyclopedia of Philosophy and the Social Sciences* (pp. 209–213). Thousand Oaks: Sage Publications Inc.

Ryle, G. (2000). *The Concept of Mind.* London: Penguin books. (Original work published 1949)

Schneider, S. (n.d.). Identity Theory. In *Internet Encyclopedia of Philosophy.* Retrieved from https://www.iep.utm.edu/identity

Schwitzgebel, E. (2019). Belief. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/fall2019/entries/belief/

Searle, J. R. (1985). Minds, Brains and Programs. In D. C. Dennett & D. R. Hofstadter (Eds.), *The Mind's I: Fantasies and reflections on self and soul* (pp. 353-373). Middlesex, England: Penguin Books. (Original work published 1980).

Searle, J. R. (1992). *The Rediscovery of the Mind.* Cambridge, MA: MIT press.

Sharples, M. et al (1994). *Computers and Thought: A practical introduction to artificial intelligence.* Cambridge, MA: MIT Press.

Shaw, M. E., & Ashton, N. (1976). Do Assembly Bonus Effects Occur on Disjunctive Tasks? A test of Steiner's theory. *Bulletin of the Psychonomic Society,* 8(6), 469-471.

Sheehy, P. (2002). On Plural Subject Theory. *Journal of Social Philosophy*, 33(3), 377-394.

Sierpinska, A. (1990). Some Remarks on Understanding in Mathematics. *For the learning of mathematics*, 10(3), 24-41.

Sierpinska, A. (1994). *Understanding in Mathematics.* London: The Falmer Press.

Skemp, R. R. (1976). Relational Understanding and Instrumental Understanding. *Mathematics teaching*, 77(1), 20-26.

Smolensky, P. (1999). Grammar-based Connectionist Approaches to Language. *Cognitive Science*, 23(4), 589-613.

Stanley, J., & Willlamson, T. (2001). Knowing How. *The Journal of Philosophy*, 98(8), 411-444.

Star, W. (2019). Counterfactuals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/fall2019/entries/counterfactuals

Steen, A., Wisniewski, M., & Benzmüller, C. (2016). Agent-based HOL Reasoning. In G.-M. Greuel, T. Koch, P. Paule, A. Sommese (Eds.), *Mathematical Software – ICMS 2016* (pp. 75-81). Berlin: Springer.

Steinberg, J. R. (2010). Dispositions and Subjunctives. *Philosophical Studies*, 148(3), 323-341.

Steiner, I. D. (1972). *Group Process and Productivity.* New York: Academic Press.

Steiner, M. (1978). Mathematical Explanation. *Philosophical Studies*, 34(2), 135-151.

Surowiecki, J. (2005). *The Wisdom of Crowds.* New York: Anchor Books.

Swart, E. (1980). The Philosophical Implications of the Four-Color Problem. *The American Mathematical Monthly,* 87(9), 697-707.

Tall, D., & Vinner, S. (1981). Concept Image and Concept Definition in Mathematics with Particular Reference to Limits and Continuity. *Educational Studies in Mathematics*, 12(2), 151-169.

Tanswell, F. S. (2017). *Proof, Rigour and Informality: A virtue account of mathematical knowledge* (Doctoral dissertation). University of St Andrews.

The Coq Proof Assistant. (n.d.). In *Coq*. Retrieved from https://coq.inria.fr/

Theiner, G. (2013). Transactive Memory Systems: A mechanistic analysis of emergent group memory. *Review of Philosophy and Psychology*, 4(1), 65-89.

Theiner, G. (2017). Group-Sized Distributed Cognitive Systems. In K. Ludwig & M. Jankovic (Eds.), *The Routledge Handbook of Collective Intentionality* (p. 233-248). New York: Routledge

Theiner, G. & O'Connor, T. (2010). The Emergence of Group Cognition. In A. Corradini & T. O'Connor (Eds.), *Emergence in Science and Philosophy* (pp. 78-117). New York: Routledge.

Theiner, G., Allen, C., & Goldstone, R. L. (2010). Recognizing Group Cognition. *Cognitive Systems Research*, 11(4), 378-395.

Thompson, P. (1998). The Nature and Role of Intuition in Mathematical Epistemology. *Philosophia*, 26(3), 279-319.

Thurston, W. P. (1998). On Proof and Progress in Mathematics. In T. Tymoczko (Ed.), In T. Tymoczko (Ed.), *New Directions in the Philosophy of Mathematics: revised and expanded* (pp. 337-355). New Jersey: Princeton University Press.

Tollefsen, D. P. (2002). Collective Intentionality and the Social Sciences. *Philosophy of the Social Sciences,* 32(1), 25-50.

Tollefsen, D. P. (2006). From Extended Mind to Collective Mind. *Cognitive Systems Research*, 7(2-3), 140-150.

Tollefsen, D. P. (2015). *Groups as Agents.* Cambridge, UK: Polity Press.

Toon, A. (2015). Where is the Understanding? *Synthese,* 192(12), 3859-3875.

Tribble, E. (2005). Distributing Cognition in the Globe. *Shakespeare Quarterly*, 56(2), 135-155.

Trout, J. D. (2002). Scientific Explanation and the Sense of Understanding. *Philosophy of Science,* 69(2), 212-233.

Trout, J. D. (2005). Paying the Price for a Theory of Explanation: de Regt's discussion of Trout. *Philosophy of Science,* 72, 198–208.

Tuomela, R. (1992). Group Beliefs. *Synthese,* 91(3), 285-318.

Tuomela, R. (2013). *Social Ontology: Collective intentionality and group agents.* New York: Oxford University Press.

Turing. A. (1985). Computing Machinery and Intelligence. In D. C. Dennett & D. R. Hofstadter (Eds.), *The Mind's I: Fantasies and reflections on self and soul* (pp. 53-67). Middlesex, England: Penguin Books. (Original work published 1950).

Tymoczko, T. (1979). The Four-Color Problem and its Philosophical Significance. *Journal of Philosophy*, 76(2), 57—83.

Van Bendegem, J. (1989). Foundations of Mathematics or Mathematical Practice: Is one forced to choose? *Philosophica,* 43, 197-213.

Van Camp, W. (2014). Explaining Understanding (or Understanding Explanation). *European Journal for Philosophy of Science*, 4(1), 95-114.

Van Kerkhove, B. & Van Bedegem, J. P. (Eds.). (2007). *Perspectives on Mathematical Practices: Bringing together philosophy of mathematics, sociology of mathematics, and mathematics education.* Dordrecht: Springer.

Varga, S. (2016). Interaction and Extended Cognition. *Synthese,* 193(8), 2469-2496. Retrieved from https://www.researchgate.net/profile/Somogy_Varga/publication/282418305_Interaction_and_extended_cognition/links/5828d70c08ae5c0137f12ea9/Interaction-and-extended-cognition.pdf

Vervloesem, K. (2007). *Computerbewijzen in de Wiskundige Praktijk* (MA Thesis). Katholieke Universiteit Leuven.

Vitale, A. S. (2017). *The End of Policing.* New York: Verso Press.

Vitale, A. S. (2020). The Police Are Not Here to Protect You. *Vice.* Retrieved from https://www.vice.com/en_uk/article/7kpvnb/end-of-policing-book-extract?utm_content=1591017564

Wegner, D. M., Giuliano, T., & Hertel, P. T. (1985). Cognitive Interdependence in Close Relationships. In W. L. Ickes (Ed.), *Compatible and incompatible relationships* (pp. 253-276). New York: Springer.

Wilkenfeld, D. A. (2013a). Understanding as Representation Manipulability. *Synthese,* 190(6), 997-1016.

Wilkenfeld, D. A. (2013b). *Explaining and understanding.* (Doctoral thesis). The Ohio State University, Columbus.

Wilkenfeld, D. A. (2017). MUDdy Understanding. *Synthese*, 194(4), 1273-1293.

Williamson, T., & Stanley, J. (2001). Knowing How. *The Journal of Philosophy*, 98(8), 411–444.

Wilson, D. S., Wilczynski, C., Wells, A., & Weiser, L. (2000). Gossip and Other Aspects of Language as Group-level Adaptations. In C. Heyes & L. Huber (Eds.), *The Evolution of Cognition* (p. 347–365). Cambridge, MA: MIT Press.

Wilson, R. A. (2004). *Boundaries of the Mind: The individual in the fragile sciences: Cognition.* New York: Cambridge University Press.

Wilson, R. A. & Foglia, L. (2015). Embodied Cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/spr2017/entries/embodied-cognition

Winograd, T., & Flores, F. (1990). *Understanding Computers and Cognition: A new foundation for design.* Norwood: Ablex.

Wittgenstein, L. (1953). *Philosophical Investigations* (G. E. M. Anscombe, Trans.). Oxford: Basil Blackwell.

Woodward, J. (2003). *Making Things Happen: A theory of causal explanation.* New York: Oxford University Press.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., Malone, T. W. (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science*, 330(6004), 686–688.

Wray, K. B. (2001). Collective Belief and Acceptance. *Synthese,* 129(3), 319-333.

Ylikoski, P. (2009). The Illusion of Depth of Understanding in Science. In H. W. de Regt, S. Leonelli & K. Eigner (Eds.), *Scientific Understanding: Philosophical Perspectives* (p.100-119). University of Pittsburgh Press.

Ylikoski, P. (2014). Agent-based Simulation and Sociological Understanding. *Perspectives on Science*, 22(3), 318-335.

Zagzebski, L. (2001). Recovering Understanding. In Steup, M. (Ed.), *Knowledge, Truth, and Duty* (pp. 235-251). New York: Oxford University Press.

Samenvatting (Nederlands)

De waarde van begrijpen [understanding] is moeilijk te ontkennen (omdat begrip een gewaardeerd doel en eigenschap is in vele activiteiten en disciplines), en de waarde van begrip filosofisch te kenmerken wordt niet langer ontkend (aangezien het concept van begrijpen zich ook heeft losgemaakt van zijn psychologische dimensie en onderscheiden wordt van het concept van verklaren [explanation] en kennis). Bijgevolg is er behoefte aan een conceptuele karakterisering die zowel filosofisch coherent en consistent als verklarend is, en die ons bovendien in staat stelt om uit te leggen wie wel en niet begrijpt, en waarom.

Ik bouw in dit proefschrift mijn conceptuele karakterisering van begrijpen en het begrijpend subject op vanuit een big picture benadering. Ik toon aan dat deze benadering een samenhangend beeld van het concept begrijpen en het begrijpend subject kan opleveren. Een beeld dat verschillende inzichten uit verschillende velden met betrekking tot de vele aspecten van het concept begrijpen kan samenbrengen, en veel van de problemen en bezwaren waartoe ze leiden kan behandelen (mede omdat ze niet behandeld moeten worden vanuit een beperkt perspectief).

Het karakteriseren van "begrip"

In het eerste hoofdstuk concentreer ik me op de markering van understanding [the mark of understanding]: welke systematische eigenschap vinden we zo filosofisch of epistemisch waardevol in het concept van begrip, en dus nodig voor zijn toeschrijving. Het concept heeft een filosofisch coherente en verklarende karakterisering nodig die consistent kan worden toegeschreven aan verschillende menselijke subjecten (en mogelijk ook niet-menselijke) met verschillende begrepen objecten en met verschillende graden van (contextuele) kwaliteit. Bovendien moet de gekozen karakterisering van het concept ons in staat stellen om om te gaan met de bekende filosofische problemen, en mogelijke tegenvoorbeelden aan te pakken.

Ik verdedig in dit hoofdstuk dat begrips-toeschrijvingen altijd verwijzen naar een bepaalde set van geschikte vaardigheden [abilities] (van een subject), samengesteld uit handelingen (relevant voor het begrepen object voor een bepaalde context), en dat dit de meest coherente en vruchtvolle conceptie is van "begrijpen." Er zijn verscheidene voordelen aan een vaardigheden-benadering. We zetten de wantrouwende rol van gevoelens opzij (zonder hen of hun rol te ontkennen). We vermijden de problemen die mentale staat [mental states] benaderingen plagen, zoals het zoeken naar een markering in een empirisch niet-waarneembaar rijk (we kunnen de mentale staat van iemand anders

niet waarnemen, en worstelen zelfs om onze eigen mentale staat adequaat te karakteriseren), diens verklarende redundantie (het is niet de mentale staat zelf die empirisch toegankelijk of epistemisch waardevol voor ons is, dus we detecteren en beoordelen een mentale staat op basis van handelingen en niet andersom) en diens vereiste voor oneindige codering (elk component van begrip zou als een staat moeten worden gecodeerd in het subject). Desalniettemin wordt er in vaardigheid-gebaseerde benaderingen ook verder gekeken dan enkel naar waarneembare handelingen (door verklarende schattingen te maken op basis van waargenomen handelingen, modaal geconceptualiseerd als contrafeitelijke [counterfactual] handelingen), maar niet naar wat erachter ligt (in een empirisch niet-waarneembaar rijk). Bovendien krijgt het concept van "impliciet begrip" [tacit understanding] meer ruimte om te bloeien (omdat epistemische vaardigheden kunnen worden gewaardeerd, zelfs zonder dat het subject er zich bewust van moet zijn), en krijgt het probleem van impliciet chauvinisme minder ruimte om te bloeien (het is moeilijker om te onderbouwen dat een bepaalde geslacht, etniciteit of zelfs soort geen begrip heeft als men een waardevol verschil in prestatie moet markeren in plaats van in een verschil in fysieke of mentale constitutie). En tot slot kunnen enkele van de nuttige concepten (die gelinkt zijn met het concept begrijpen) die niet duidelijk passen bij een op vaardigheid-gebaseerde benadering (bijv. de overtuigingen [beliefs] van subjecten en de betekenis [meaning] van objecten) niet alleen hun verklarende kracht behouden, maar ze zijn nog sterker geworteld als instrumentele concepten die zijn afgeleid van handelingen.

Nadat ik het "begrijpen" heb gekarakteriseerd als vaardigheden, beschouw ik in het kort enkele soorten vaardigheden die door de literatuur worden aangeboden als de juiste. Ik zal aangeven dat ik deze niet als een noodzakelijke of voldoende voorwaarde [necessary or sufficient condition] voor het begrijpen zal beschouwen, maar als dat wat de reikwijdte van understanding vormt.

Het karakteriseren van de kwaliteit

In het tweede hoofdstuk conceptualiseer ik de dimensies en graden van kwaliteit in begrijpen, bied ik een contextuele benadering aan om te kunnen specificeren wat er relevant is, en specificeer ik enkele van de problemen en kansen waartoe dit leidt voor het evalueren van begrip. Ook al erkennen de meeste auteurs de graden van understanding, behandelen maar weinigen hen zo expliciet als hier. We zien vier dimensies van kwaliteit, waarvan er drie waren samengesteld uit twee parameters, één die het verbreedt en één die het verdiept (zie samenvatting hieronder). Helaas, en niet geheel verrassend, kan geen enkele overeengekomen universele standaard alle toeschrijvingen van understanding binnen deze dimensies verduidelijken. We vinden niet alleen contextuele variatie in drempels, maar ook in wat in de eerste plaats als relevant wordt beschouwd. Daarom bied ik ook een

contextuele benadering aan binnen elk van deze dimensies en parameters, zodat elk van hen kan variëren naarmate wat relevant of passend is (zie samenvatting hieronder). De verantwoording van wat passend is laat ik over aan een andere discussie.

Dimensies	*Parameters*	*Contextuele specificering*
Reikwijdte [Scope] van vaardigheden	*Reikwijdte* de hoeveelheid variatie in capaciteiten die een rol spelen bij het begrijpen van een object X.	welke capaciteiten die verbonden zijn met X als saillant worden beschouwd, en in welke mate.
Gevoeligheid [Sensitivity] van een vaardigheid	*Situationeel reactievermogen [Situational responsiveness]* geschikte veranderingen in de prestaties op veranderingen in de object-situatie, bijv. Reageren op wat-als.	welke variaties in object-situaties, samen met hun gepaste reacties, relevant zijn, en in welke mate.
	Nauwkeurigheidsgraad [Accuracy] nauwkeurigheid in prestatie, bijv. Aantal decimalen.	welke soorten nauwkeurigheid relevant zijn (waar toepasselijk) en in welke mate.
Stabiliteit [Stability] van een handeling	*Bereik [Range]* aanwezigheid in (contra-)feitelijke omstandigheden.	welke soorten (contra-)feitelijke omstandigheden, waarbij hetzelfde epistemische subject naar behoren handelt, relevant zijn en in welke mate.
	Robuustheid [Robustness] aanwezigheid na (contra-)feitelijke omstandigheden.	welke soorten omstandigheden, waarna het onderwerp op de juiste manier moet blijven handelen, opvallend zijn en in welke mate.
Efficiëntie [Efficacy] tegenover een onderwerp	*Economie [Economy]* de juiste handeling gebruikt een minimum aan relevante toegestane middelen.	welke specifieke middelen (incl. Evenementen) men bereid is toe te staan om te worden gebruikt om de verworven vermogens te beschouwen, en in welke mate.
	Potentieel [Potential] de juiste handeling wordt verkregen met de toevoeging van een minimum aan relevante bronnen of evenementen.	welke specifieke middelen en omstandigheden men bereid is in overweging te nemen om de ermee bereikte capaciteiten te beoordelen, en in welke mate.

Ik beargumenteer dat de meeste toeschrijvingen van begrip neerkomen op een bewering over de graad binnen deze dimensies. Zelfs als deze parameters niet perfect zijn om een "ideaal" of kwantitatieve beoordeling van de kwaliteit van het begrijpen te conceptualiseren, zijn ze vruchtbaar in het diagnosticeren van de sterke en zwakke punten, de soorten en verschillen in kwaliteit, evenals de problemen of kansen bij de evaluatie (bijv. “kludges”, indirect versus direct bewijs, soorten begrip en maximaal begrip).

Het aanpakken van (tegen)voorbeelden

In Hoofdstuk 3 krijgen we blijk van hoe sterk mijn contextuele vaardigheids-benadering het doet in het aanpakken van lastige of misleidende toeschrijvingen, vaak voorkomende voorbeelden en

mogelijke tegenvoorbeelden. Hier leg ik uit dat de meeste voorbeelden van begrijpen zonder vaardigheden (via handelingen) uiteindelijk toch worden verantwoordt via vaardige handelingen. Dit kan ik doen aan de hand van contextuele variatie en een contrafeitelijke kijk op het begrip vaardigheid. Vervolgens kan ik ook aantonen dat voorbeelden van vaardigheden zonder begrip uiteindelijk worden verantwoord via een beperking in handelingen (te conceptualiseren via de parameters uit Hoofdstuk 2), of door zich op het verkeerde subject te richten. Hiermee maak ik aanstalten om het subject nader te bekijken in het volgende hoofdstuk.

Het karakteriseren van het subject

Tot heden toe wordt vaak verondersteld dat het relevante subject waaraan begrip kan worden toegeschreven altijd menselijke individuen zijn (of zouden moeten zijn). De markering van het begrijpend subject [the mark of the understanding subject] is niet met even grote zorg behandeld door de (epistemologische) literatuur als de markering van het begrip zelf. Desalniettemin zijn er veel gevallen die deze veronderstelling betwisten, en we hebben een manier nodig om dit met consistentie en zonder antropocentrische vooringenomenheid te conceptualiseren. Veel van de vaardigheden die we tegenwoordig in de wetenschap en in het dagelijks leven vinden, worden steeds vaker bekrachtigd door entiteiten voorbij aan (enkel) menselijke individuen. Vaardigheden worden bekrachtigd met behulp van instrumenten (bijv. Pen en papier om een proef uit te werken, een kalender om een stap in een uitgebreid experiment te onthouden, een interactieve theorema-bewijzer [interactive theorem prover] om nieuwe wiskundige bewijzen te ontdekken, en veel andere apps, enz.), door een groep individuen (bijv. via een disjunctieve verdeling van antwoorden in een quiz team, een conjunctieve/ additieve datapool in een onderzoekscentrum, een gemiddelde-functie van schattingen in een menigte, een coöperatieve arbeidsverdeling in een laboratoriumexperiment of een dynamische interactie in een koppel) of zelfs door kunstmatige systemen (bijv. navigatiesystemen, geautomatiseerde programma's die op zichzelf natuurwetten of wiskundige bewijzen afleiden, of software die verkeersstroom, regen of storm voorspelt). Wat we dus nodig hebben is een markering van het epistemisch subject [mark of epistemic subjecthood] dat ons kan helpen om een relatief samenhangende, coherente en aanhoudende entiteit aan te duiden waarvoor toeschrijvingen van epistemische eigenschappen (bijvoorbeeld begrip, kennis, overtuigingen, enz.) verklarend of voorspellend kunnen zijn.

Als markering van het epistemisch subject heb ik de interpretatieve [interpretationist] benadering verdedigd, in het bijzonder de “epistemic stance” (de “intentional stance” met een epistemische focus). De “epistemic stance” is de strategie om gedrag te te interpreteren alsof het een epistemische

agent [epistemic agent] is die wordt beheerst door overtuigingen [beliefs], epistemische doelen [epistemic aims] (d.w.z. het soort resultaten dat een epistemische praktijk waardeert) en epistemische tactieken [epistemic tactics] (d.w.z. elke serieuze, systematische poging om dichter bij een epistemisch resultaat te komen). Het is essentieel dat de enige verantwoording voor het mogen interpreteren van een entiteit als epistemisch agent te vinden is in het verklarende en voorspellende succes van die interpretatie. De "epistemic stance" stelt ons in staat om de innerlijke werking (bijv. overtuigingen, doelen, tactieken) te conceptualiseren op basis van handelingen, en ons toestaat om de agent af te bakenen op basis van deze interpretatie (door te zoeken naar de realiserende basis voor de gepostuleerde epistemische agent).

Uitgebreide begrijpers

Nadat ik de "epistemic stance" heb verdedigd als de markering van het epistemisch subject, zal ik Hoofdstuk 4 afsluiten met te argumenteren dat, als de "epistemic stance" tegenover een uitgebreide entiteit (bestaande uit meer dan alleen een menselijk individu - bijv. een mens met een notaboekje) ons verklarende of voorspellende krachten kan geven, dan is het profiteren van deze kracht niet alleen verantwoord en vruchtbaar, maar ook consistent met onze beste conceptualisaties van menselijke individuen. In dat geval hebben we te maken met een uitgebreid, epistemisch agent [extended epistemic agent]. Uitgebreid begrip [extended understanding] betekent dat de realiserende basis van het begrijpend subject zich verder uitstrekt dan een menselijk individu alleen. Dit omvat altijd een uitgebreide realiserende basis, maar kan ook andere uitbreidingen omvatten, zoals de uitbreiding van de handelende basis of de uitbreiding het epistemisch agentschap. Op het einde van Hoofdstuk 4 bespreek ik deze onderscheidingen aan de hand van 7 voorbeelden.

Collectieve begrijpers

In het vijfde hoofdstuk richt ik mijn aandacht op collectief begrip [collective understanding]. Er zijn talloze voorbeelden in natuurlijke taal waarin een groep van mensen samen met begrip worden toegeschreven. Zijn deze toeschrijvingen verondersteld louter lege retoriek, overbodige metaforen en handige afkortingen te zijn, of schuilt er een diepere verklarende kracht achter? Hoewel ik niet afdoende zal kunnen antwoorden of kandidaten van epistemische groep agenten [epistemic group agents] bestaan, zal ik licht werpen op de conceptuele ruimte die betrokken is bij het onderbouwen van een dergelijk antwoord. Ik zal beweren dat een aantal basisstappen voor een groep moeten worden doorlopen om het toeschrijven van "collectief begrip" te verantwoorden. Eerst en vooral moet er een groep zijn. Ten tweede moet die groep vaardigheden vertonen, als groep. En ten derde moeten die vaardigheden resulteren in een succesvolle "epistemic stance."

Maar zelfs als een groep menselijke individuen zich als één geheel gedraagt (en zo een verklarend kracht geeft aan de “epistemic stance”), is het nog steeds mogelijk om de verklaring op groepsniveau te reduceren tot een verklaring op individueel niveau - waardoor de verklaring op groepsniveau overbodig wordt. Dit is het reduceerbaarheidsprobleem [the reducibility problem]. Wanneer is zo'n reduceerbaarheid een probleem en wanneer niet? Ik zal betogen dat reduceerbaarheid een probleem is wanneer de vaardigheden en het epistemisch agentschap [epistemic agency] van de groep systematisch ge”mapped” kunnen worden op een conglomeraat van die van haar leden. De laatste stap kan dus enkel verwezenlijkt worden als er geen dergelijke “mapping”-relatie te vinden is door de het emergerende [emergent] “assembly bonus effect.” Ik zal hiervan zowel een ideaal hypothetisch voorbeeld geven als een korte uiteenzetting van twee bestaande voorbeelden.

Artificiële begrijpers

In het zesde en laatste hoofdstuk schenk ik mijn aandacht aan kunstmatig begrip [artificial understanding]. Kunnen kunstmatige systemen (zoals computers) ooit met begrip worden toegeschreven? Om deze vraag positief te beantwoorden moet eerst worden vastgesteld of kunstmatige systemen epistemische vaardigheden kunnen vertonen en of dergelijke vaardigheden tot een succesvolle “epistemic stance” kunnen leiden. Dit wordt gemakkelijk aangetoond.

Er zijn echter enkele argumenten die blijken aan te tonen dat “kunstmatig begrip” in principe onmogelijk is, (zelf in de aanwezigheid van vaardigheden of het succes van de “epistemic stance”). Ze hebben betrekking op het probleem van regressie [the regress problem], reduceerbaarheid [the reducibility problem] en rigiditeit [the rigidity problem]. Deze vormen conceptuele hindernissen die we moeten overwinnen om de in principe mogelijkheid van kunstmatig begrip te kunnen verantwoorden. Het overwinnen van de eerste twee conceptuele hindernissen houdt in dat je een antwoord hebt op de vraag: waarom zijn de vermeende vaardigheden of het epistemisch agentschap van een kunstmatig systeem (en dus diens begrip) niet te regresseren naar diens programmeur of reduceren tot diens programmering? Als je de vaardigheden of epistemische eigenschappen van het kunstmatige systeem eenvoudig kunt “mappen” op die van zijn programmeur of aan de procedures in zijn programmering is het overduidelijk overbodig om een extra agent te postuleren en de vaardigheden toe te schrijven aan het systeem als geheel. Maar het probleem van regressie en reduceerbaarheid neemt de legitieme kritiek op een overbodige epistemische agent aan en breidt deze ten onrechte uit naar elk geval waar er een oorzakelijke oorsprong of superveniëntie

[supervenience] is, ongeacht hoe handig, zelfvoorzienend of verklarend krachtig het is om de entiteit op zichzelf te beschouwen.

Om de derde conceptuele hindernis te overwinnen, moet je kunnen antwoorden waarom kunstmatige systemen niet te rigide zijn om de volledige reikwijdte van vaardigheden die we bij mensen vinden weer te kunnen geven. Ik zal betogen dat het rigiditeitsprobleem er ten onrechte van uitgaat dat het computationeel niveau [computational level] moet overeenstemmen met het niveau van vaardigheden, terwijl emergentie [emergence] ervoor zorgt dat het computationele niveau ver onder dat niveau kunnen liggen.