Extending Cognition in Epistemology

Towards an Individualistic Social Epistemology

Spyridon Orestis Palermos

PhD by Research in Philosophy
The University of Edinburgh
2012

To Elvira, Vasso, Dionysis, and in memory of Spyros;
my beloved grandparents.

Table of Contents

Part 1

Part 2

Declaration

I, Spyridon Orestis Palermos, hereby declare the following. The present thesis, submitted for examination in pursuit of a PhD by Research in Philosophy, has been entirely composed by myself, and it has not been submitted in pursuit of any other academic degree, or professional qualification.

Signature:

Date:

Acknowledgements

Those familiar with the academic world of philosophy will probably note that this is an 'Edinburgh' thesis. I couldn't agree more; indeed, I cannot imagine undertaking this work in any other institution. But I must also admit I could have never arrived in such a friendly and stimulating environment without the support of loving and trusting people life has blessed me with. To avoid, however, being (too) personal, I will try to thank those who immediately come in mind, and do so in categories—although many of them should probably show up in more than one place.

First of all, I cannot express enough gratitude to my supervisors, Duncan Pritchard, Mark Sprevak, and Andy Clark. Their philosophical work has being an immense inspiration, and in its absence I could have produced hardly any arguments for my thesis. Moreover, they have all proved to be excellent advisors. They have spent ample time providing me both with professional advice and invaluable feedback on my work. And they have all treated me in the wisest way, allowing for me to genuinely enjoy working on my thesis.

Also, I am largely indebted to the faculty and postgraduate community of the philosophy department at the University of Edinburgh, in general. They afforded a stimulating intellectual circle in which I could develop and try out my ideas, and they helped me gain confidence in my academic pursuit. Some of the people I can single out are Shane Ryan and Eusebio Waweru, with whom I have had many interesting conversations. Also, exchanging opinions and literature with Evan Butts (who received his PhD on a similar topic to mine, last year) has been distinctively helpful. Above all, however, I owe a sincere debt of gratitude to Barbara Scholz who unexpectedly passed away a year ago. Barbara was one of my first philosophy teachers. I met her when I first arrived in Edinburgh for my MSc degree, and she was the first person who encouraged me to apply for a PhD in philosophy.

Apart from my academic friends, however, I must also thank some people who have been personally close to me. One way or another, they

helped me reach here and kept me going. First, I owe a great deal to my dad for nudging me to carry on with my (long) engineering studies ("You never know", he vaguely used to say), reassuring me that if I so wished, I could then convert to anything ("even philosophy"). An advice that has already paid off—many of my arguments rely on what I was taught back then. I also thank him for having always surrounded me with many inspiring books; although I have to thank my mom—my favorite teacher—for helping me overcome my teen-denial and actually go on and read (some of) them. In addition, they both showed a huge amount of trust in me and helped me make the choice to switch to philosophy. And then, I cannot help mentioning my brother, Dennis, for showing a genuine interest in philosophy and in what I have been working on all these years, despite having just turned 17. Probably, he is the only one who might try reading some of it.

Next are my two best Edinburgh friends, Sara De Las Heras, and Scott Kerzman, who have been incredibly considerate and supportive. In addition to that, they are also responsible for all those nights of "unbridled debauchery" (Scott's words) we spent together, and which I deem necessary for dealing with a PhD degree. Scott is also a fellow colleague and, during many of those nights, we had several passionate and productive debates.

And finally, having commented on how my dissertation is specially related to its place of origin, I need to thank the two people that helped me arrive here. Four years ago, I knew nothing about academic philosophy. In fact, it was Niki Andrikopoulou who suggested the University of Edinburgh and agreed to accompany me here; she would study poetry, while I would study philosophy—and so it happened. If it hadn't been for her, I wouldn't have found my way to this destination. And the person who helped me make my way here is my first philosophy mentor, Professor Aristides Baltas. He introduced me to the world of philosophy, radically changed my understanding of science, and guided me through a transition that I could have never made on my own.

My sincere thanks to all of you.

ABSTRACT

The aim of the present thesis is to reconcile two opposing intuitions; one originating from mainstream individualistic epistemology and the other one from social epistemology. In particular, conceiving of knowledge as a cognitive phenomenon, mainstream epistemologists focus on the individual as the proper epistemic subject. Yet, clearly, knowledge-acquisition many times appears to be a social process and, sometimes, to such an extent—as in the case of scientific knowledge—that it has been argued there might be knowledge that is not possessed by *any* individual alone. In order to make sense of such contradictory claims, I combine virtue reliabilism in mainstream epistemology with two hypotheses from externalist philosophy of mind, *viz.*, the extended and distributed cognition hypotheses. Reading virtue reliabilism along the lines suggested by the hypothesis of extended cognition allows for a weak anti-individualistic understanding of knowledge, which has already been suggested on the basis of considerations about testimonial knowledge: knowledge, many times, has a dual nature; it is both social and individual. Provided, however, the possibility of distributed cognition and group agency, we can go even further by making a case for a robust version of anti-individualism in mainstream epistemology. This is because knowledge may not always be the product of any individual's cognitive ability and, thereby, not creditable to any individual alone. Knowledge, instead, might be the product of an epistemic group agent's collective cognitive ability and, thus, attributable only to the group as a whole. Still, however, being able—on the basis of the hypothesis of distributed cognition—to recognize a group as a cognitive subject in itself allows for proponents of virtue reliabilism to legitimately apply their individualistic theory of knowledge to such extreme cases as well. Put another way, mainstream individualistic epistemologists now have the means to make sense of the claim that *p* is known by *S*, even though it is not known by any individual alone.

FOREWORD

The driving force behind the present dissertation is the intuition that much of our knowledge is social. This is not necessarily to deny that it is also individual—although, as we shall see, this denial might, in certain cases, be correct as well.

On the contrary, mainstream epistemology has traditionally been individualistic, thereby obscuring from view and suppressing the social nature of knowledge. That is, conceiving of knowledge as a cognitive phenomenon, mainstream epistemologists have tended to focus on the individual as the proper epistemic subject. Cognition, after all, it is largely held, rests within the individual's head. Accordingly, if one is to account for knowledge, one should focus on the cognitive/epistemic properties of the individual agent. So, having such methodological considerations in mind, makes it unsurprising that, until recently, the most popular epistemological approach was that of epistemic internalism: One knows some true proposition only if one has, in principle, internal access (i.e., by reflection alone) to the reasons/justification for one's true belief in that proposition. And such a view, in turn, appears to entail the demand for intellectual autonomy. If one must be internally justified for one's true belief, then one's reasons for holding that belief cannot originate from anywhere else, but from oneself, alone. Hence, knowledge cannot be social.

Surely, however, not all knowledge can be like that. Testimonial knowledge, for instance, appears to be a clear counterexample, as it depends both on the hearer's and the speaker's reasons for believing the reported proposition. In other words, knowledge acquisition many times appears to be a social process. And sometimes—as in the case of scientific knowledge—to such an extent that it has been argued (Hardwig 1985) there might be knowledge that is not possessed by *any* individual alone. Accordingly, such and similar considerations have given rise to social epistemology, which, reasonably, is wildly held to be at odds with individualistic epistemology.

Here is then, stated in general terms, the main question my thesis aims to address: How can we bridge the unsettling gap between mainstream individualistic epistemology and social epistemology—how can those two distinct theoretical domains with the same subject matter be brought together? In other words, could there be one single account of knowledge able to deal not just with individual knowledge, but also with knowledge that is partly individual and partly social, or even with knowledge that is entirely social?

In order to provide a positive response to the above questions I will employ considerations that originate from recent advances within philosophy of mind and cognitive science. In particular, I will focus on two distinct, yet interrelated, hypotheses within externalist philosophy of mind, namely the extended and distributed cognition hypotheses (HEC and HDC, respectively). Both of them go against traditional internalist philosophy of mind, which holds that cognition is restricted within the individual's head, or at most her organismic boundaries (reminiscent to epistemic internalism that holds that one's justification should be restricted within one's head). The first one does so by claiming that cognition extends to the epistemic artifacts an agent might employ, while the second goes even further by postulating that cognition might be distributed across a group of individuals and their epistemic artifacts. Both of these theses are the extreme consequents of a paradigm shift within cognitive science, namely the approach of embodied and embedded cognition. This was the result of the recognition that an agent's brain-functions are heavily dependent on the agent's body and environment. Several embodied cognition theorists, however, have claimed that an agent's brain and his/her body are interdependent to such an extent that we should consider the agent's body as a constitutive element of the agent's overall cognitive system (notice that this move is sometimes allowed by opponents of HEC and HDC). HEC and HDC theorists, however, go even further by claiming that when an agent's internal cognitive processes and specific aspects of his/her environment (these aspects can be epistemic artifacts such as pen and paper, or other individuals) are heavily interdependent then cognition is not just embedded. Instead, in such cases, there is an overall extended, or distributed cognitive system that comprises of both the agent and his/her epistemic artifacts, or the individuals he/she is mutually interacting with.

But how can such hypotheses be associated with mainstream epistemology? Fortunately not all mainstream epistemology is internalist. As mentioned before, this has been so only until recently. In the second half of the 20th century, due to independently motivated reasons (mainly having to do with Gettier's (1963) counterexamples to epistemic internalism and the problems of Humean and radical skepticism) there appeared epistemic externalism. Epistemic externalism is the denial of epistemic internalism. That is, according to externalists in epistemology, the knower does not need to have internal access to the reasons of holding her true beliefs. So long as one's true beliefs are reliably and/or safely formed then one can be said to know even if one has no beliefs about the reasons for one's true beliefs. Now, notice that in so denying the demand for internal access to the reasons for holding a true belief seriously undermines the accompanying commitment to intellectual autonomy and, thus, opens the way for anti-individualist (i.e., social) epistemology.

I will here concentrate on a kind of externalist epistemology that only commits itself to the ability intuition on knowledge. This is the idea that knowledge must be the product of cognitive ability. Crucially, however, virtue reliabilism (VR)—as the target view is known in the literature—makes no claims about what may count as a cognitive ability, thereby being open to interpretations along the lines suggested by HEC and HDC. In particular, VR, roughly, is the view that knowledge is creditable true believing that is true in virtue of one's cognitive ability. This is the starting point of my thesis, which is divided in two parts. Part 1 (chapters 1, 2, and 3) presents the broader virtue reliabilistic framework and HEC, and demonstrates how the two views can be combined. Part 2 (chapters 4, 5, and 6) explores the ramifications of reading virtue reliabilism along the lines suggested by HEC and HDC, and demonstrates how such a reading can reveal and account for the social nature of several kinds of knowledge.

In more detail, in chapter 1, I motivate VR on the basis of Humean skepticism and several thought experiments, all of which will allow us to delineate the core tenets of the view. As I further argue, however, virtue reliabilism, as it stands, cannot account for certain crucial counterexamples (including testimonial knowledge), and it cannot explain how one can have knowledge that is the product of the operation of epistemic artifacts. Only after we embrace a virtue reliabilistic necessary condition on knowledge that has been recently proposed by Pritchard—namely, COGA$_{weak}$—can we account for testimonial knowledge. Moreover, on the basis of COGA$_{weak}$ we can also claim that the operation of some epistemic artifact can count as a *bona fide* cognitive ability, such that we can acquire knowledge on its basis, while being in line with the ability intuition on knowledge.

Claiming, however, that epistemic artifacts can be part of one's cognitive system is a rather radical claim, which had better not been left metaphysically unsupported. Accordingly, in chapter 2, I explore HEC. In particular, I offer a detailed defense of the view from its most serious objections—namely the 'coupling-constitution' fallacy and 'cognitive bloat' worry—on the basis of Dynamical Systems Theory (DST). The outcome is a set of necessary and jointly sufficient conditions on cognitive extension that can safeguard HEC from the aforementioned objections, as well as clearly distinguish it from the hypothesis of embedded cognition.

Interestingly, turning back to COGA$_{weak}$, in chapter 3, I argue that the very same set of criteria is required for a process to be knowledge-conducive. And this is as it should be; COGA$_{weak}$ holds that knowledge must be the product of cognitive ability and HEC sets out to reveal which processes can count as cognitive abilities without being distracted by the arbitrary boundaries of skin and skull.

Having so wedded the two views—which is the aim of Part 1—I turn to the second part, which is dedicated to revealing the social nature of much of our knowledge—mainly by focusing on the ramifications of reading COGA$_{weak}$ along the lines suggested by HEC and HDC. First, in chapter 4, I am interested in the claim that many instances of knowledge have a dual nature—they are both social and individual. Focusing, for example, on testimonial knowledge and knowledge acquired on the basis of epistemic artifacts—both software and hardware—these are true beliefs that are creditable both to the individual who appropriately accepted the offered report, or employed the relevant artifacts, *and* to the individual(s) who offered the reliable report, or brought the relevant artifacts about (remember that according to VR and COGA$_{weak}$, knowledge is *creditable* true believing, which is true in virtue of one's cognitive ability). We therefore now have a good case for what might be called a weak version of anti-individualism in mainstream epistemology.

Provided, however, the possibility of distributed cognition and group agency, we can go even further by making a case for a robust version of anti-individualism in epistemology. This is the theme of the last two chapters. In chapter 5, I motivate the existence of group agents by combining the consequences of the phenomena of multiple realizability and wild disjunction for constructing scientific laws with the arguments from DST that I used to argue for HEC, in chapter 2. Having so argued for the existence of group agents, I then ask whether there could be *epistemic* group agents.

And considering Transactive Memory Systems and scientific experiments performed by research teams, I argue in chapter 6, points to a positive answer. Both cases appear to be about epistemic group agents, which exist and gain knowledge in virtue of non-reducible collective belief-forming processes. The interesting point, however, is that this is knowledge that is not produced by any individual's cognitive ability and, thereby, not creditable to any individual alone. Instead, in such cases, knowledge is the product of the group's collective cognitive ability and must be attributed to the group as a whole. Still, however, being able to recognize a group as a cognitive subject in itself, allows for the proponents of VR, and $\text{COGA}_{\text{weak}}$ in particular, to legitimately apply their individualistic theory of knowledge to such extreme cases as well.

In other words, by the end, we will have a picture according to which we can use a mainstream individualistic approach to knowledge that can account for all kinds of knowledge; i.e., knowledge that is strictly individual, knowledge that is partly individual and partly social, and knowledge that is entirely social (i.e., not possessed by any individual alone). And if such a picture is possible, then the prospects for a unified epistemology that goes beyond the boarders of individualistic and social epistemology should be rich.

Finally, before closing this foreword, let me also stress two points regarding the dialectics of the thesis. First, for the sake of intellectual honesty, let me note a worry that I, myself, have with respect to the metaphysical support I offer to HEC and HDC in chapters 2 and 5, respectively. In both cases, I use arguments from DST, and in the case of HDC I further combine them with considerations about what might count as a lawful causal explanation within science. Both argumentative lines proceed this way: Given how science is performed and/or DST (the best available mathematical language for the study of complex systems) we must accept that HDC and HEC are correct, respectively. So both arguments rely on what may count as a good explanation, and given that explanatory concerns are epistemological concerns, it now appears that neither of these two argumentative lines is strictly metaphysical, i.e., independent of what might one say or know. Yet it is also true that much of the metaphysics within contemporary philosophy of mind is done this way: so long as X does significant explanatory work then X cannot be metaphysically eliminated and should, thereby, be considered as real. I leave it to the reader to decide what this skeptical point may amount to.

Second, I want to warn the reader about my attitude towards the several theses I employ for the sake of my arguments. As you will find out I write as if I completely buy into all COGA$_{weak}$, HEC, and HDC. This is partly because I really do. I recognize, however, that I could have resisted this risky and maybe provocative attitude. Perhaps, I could have instead presented the views, thoroughly criticized them, demonstrated that none of the criticisms is conclusive, and further argued for what happens if we combine them. I am afraid, however, that such an approach would have resulted in a much longer manuscript, including details that would be irrelevant to my main thesis. Moreover, I find it hard to see how one could present a positive picture by being overly defensive. Instead, I have preferred to offer what I consider to be the most serious objections facing the views I support, and provide answers to them, arguing, in effect, that one could non-problematically cling to them. Since, however, some of these theses are widely thought to be radical, or even implausible, I offer this additional warning note, and I leave it to the reader to form an opinion about their force on the basis of his/her intuitions and the arguments offered in the pages to follow.

PART 1

CHAPTER 1

Virtue Reliabilism and COGA$_{weak}$

1.1) Introduction

In this first chapter the focus will be on the introduction of a necessary condition on knowledge that has been recently put forward by Pritchard (2010*b*), namely COGA$_{\text{weak}}$. COGA$_{\text{weak}}$ describes knowledge as a kind of action which must be significantly creditable to one's cognitive agency for acting on her reliable belief-forming dispositions in order to get things right. The reason I am interested in this particular condition is twofold. Apart from fitting nicely with the Hypothesis of Extended Cognition (HEC) within contemporary philosophy of mind (more on this fit in chapter 3), it also appears to be a (if not the most) promising formulation of a necessary condition on knowledge, which is capable to accommodate a wide range of diverse epistemological considerations.

The last claim will become gradually obvious through the discussion of how COGA$_{\text{weak}}$ captures a fundamental insight concerning the nature of knowledge. I am referring to the *ability intuition* on knowledge, which expresses the popular, amongst contemporary epistemologists, realization that knowledge must be the product of cognitive ability. To get a grip on the importance of this guiding idea, we will first concentrate on process reliabilism and, then, we will move on to the consideration of a subsequent proposal, that of virtue reliabilism. By then, the background will be clear enough for introducing COGA$_{\text{weak}}$.

1.2) Process Reliabilism

COGA$_{\text{weak}}$, as we shall see later on, is a virtue reliabilistic necessary condition on knowledge. By way of introducing it, therefore, we must first talk both about virtue reliabilism and process reliabilism, of which the former is a descendant view. Process reliabilism is an externalist approach to knowledge in that it denies that the agent must have—even in principle—reflective access to the reasons for which his beliefs are true. In order to understand and appreciate the reason for this denial let me briefly note what is thought to be one of the primary motivations for it.

If we demanded that one always have, at least in principle, reflective access to the reasons for which one's beliefs are true—as internalist theories of knowledge do—then one would need to be able to provide deductive arguments as reasons for holding one's beliefs about unobserved matters of fact and the external world. As Hume's skeptical arguments demonstrate, however, this is impossible. Accordingly, it has been traditionally assumed that Hume's arguments lead to skepticism about our empirical beliefs.

The problem of induction is well known, but let me expand on it a bit. We form our beliefs about unobserved matters of facts and the external world on the basis of evidence provided by past and present observations, and the external world, respectively. But, in order for our empirical conclusions to logically (i.e., necessarily) follow from the evidence offered in their support, we must also make the assumptions that the future will resemble the past and that sensory appearances are a reliable indication to reality, respectively. The problem, however, is that both of these assumptions rely for their support on what they assert. So, given that circular reasoning cannot give rise to knowledge, they cannot be known. Consequently, since all our empirical beliefs depend for their justification on unknown assumptions, they cannot logically (i.e., necessarily) follow from the evidence offered in their support. Accordingly, the conclusion that has been traditionally drawn is that our empirical beliefs cannot amount to knowledge.

Contemplating on Hume's skeptical arguments, however, Greco (1999) argues that this is too fast. Hume's arguments should not be considered as one-way skeptical ones. Instead, the immediate conclusion to be drawn from them is that our empirical beliefs do not necessarily follow from their evidence; if the evidence for our empirical beliefs is reliable then it is at most contingently reliable. This realization alone, however, cannot automatically lead to skepticism. Only after we embrace an internalist understanding of knowledge, such that one's beliefs should always necessarily follow from their evidence, do we face skepticism.

In other words, in order to avoid skepticism about empirical and perceptual knowledge, we must allow knowledge to be grounded on evidence that is merely contingently reliable, and so we must give up the requirement that one's beliefs should always be internally—i.e., by reflection alone—justified. Put another way, we must recognize that if the evidence offered in favor of our empirical and perceptual beliefs is indeed a reliable indication of the truth—thereby delivering knowledge—then "this is at most a contingent fact about human cognition, rather than a function of any necessary relations, deductive or inductive, between evidence and belief" (Greco 1999, 273). Accordingly, argues Greco, any adequate epistemology must be able to account for the fact that *merely contingently reliable evidence can give rise to knowledge* (1999, 273). Are there, however, any resources for supporting this radical alternative?

At this juncture, Greco puts forward process reliabilism, which is the idea that knowledge is true belief that is the product of reliable belief-forming processes, where a reliable process is a process that tends to result in true rather than false beliefs. Moreover, in response to our skeptical considerations, "reliabilism denies that evidential relations must be necessary, and denies that one must know that one's evidence is reliable" (Greco 1999, 284); if forming a belief on a certain kind of evidence constitutes a reliable belief-forming process, it does not matter that the evidence is only contingently reliable. "Reliabilism makes *de facto* reliability the grounds of positive epistemic status" (Greco 1999, 284-5).

Notice, then, that process reliabilism is an externalist approach to the theory of knowledge. On this view—contrary to the traditional account of knowledge as internally justified true belief—in order to know one does not need to know, or be justified in believing (by reflection alone, or any other means) that one's beliefs are formed in a reliable fashion. So long as one employs a reliable belief-forming process one is justified in holding the resulting beliefs.[1]

[1] Notice, however, that Greco concedes that a plausible account of knowledge should be able to satisfy the intuition that one must also be sensitive to the reliability of one's evidence. But, if, as Hume's skeptical reasoning demonstrates, the relation between evidence and belief is not necessary, then it is far from obvious how a person can be so subjectively sensitive, especially in externalist approaches such as process reliabilism; if a condition of 'subjective sensitivity to the reliability of one's evidence' must be satisfied then this should better be accomplished in a way that will not require *knowledge of, or even beliefs about* the said reliability. Otherwise, at least in cases of empirical knowledge, such a requirement would

So we see that process reliabilism has the resources to overcome the Humean skepticism. There is, however, a serious complication with the view. That is, process reliabilism, as it stands, is too weak a condition on knowledge because it allows *any* reliable belief-forming process to count as knowledge-conducive, and this is intuitively incorrect. Consider the following three examples.

> Hercules (Adapted from Pritchard's Temp (2009, 48))
>
> Hercules tosses a drachma whenever he wants to form a belief about the weather outside. If it is heads, he forms the belief that it is sunny; if it is tails he believes it is cloudy; and if it balances in between, he believes it is rainy. As it happens, Hercules' way of forming his weather beliefs is perfectly reliable, because Zeus, who wants to save Hercules from the embarrassment of forming false weather beliefs, has an eye on him; every time he sees Hercules tossing the coin arranges the world accordingly.

> Serendipitous Brain Lesion (Greco 2010, 149)
>
> Suppose that *S* has a rare brain lesion, one effect of which is to reliably cause the true belief that one has a brain lesion. Even if the process is perfectly reliable, it seems wrong that one can come to have knowledge that one has a brain lesion on this basis.

> Careless Math Student (Greco 2010, 149)
>
> Suppose that *S* is taking a math test and adopts a correct algorithm for solving a problem. But suppose that *S* has no understanding that the algorithm is the correct one to use for this problem. Rather, *S* chooses it on a whim, but could just as well have chosen one that is incorrect. By hypothesis, the algorithm is the right one, and so using it to solve the problem constitutes a reliable process. It seems wrong to say that *S* thereby knows the answer to the problem, however.

How does process reliabilism rule in these cases? Hercules' beliefs are formed in a highly reliable way and so, on the adjudicated view, Hercules has knowledge of the weather conditions. Intuitively, however, knowledge should not be attributed to Hercules. Why not? On a first analysis, we may just say that in cases of knowledge we want our beliefs to be responsive to the

drive us straight back to the Humean problematic. In section 1.3.2, we will see how Greco accommodates this intuition—which has arguably been the central motivation for internalist theories of knowledge—but in a way that avoids the Humean problematic.

facts, whereas in Hercules' case, the direction of fit is exactly the opposite; it is not Hercules' beliefs that agree with facts, but the other way around. Likewise, the unfortunate agent's way of forming his true belief about his brain lesion on the basis of his brain lesion is reliable as well. Accordingly, process reliabilists must accept that he can gain knowledge in this way. As, Greco claims, however, this doesn't sound correct. Why not? Mainly, because the way the agent forms his belief is so strange that no one could accept that one can gain knowledge in this way—not even the agent himself. And finally, the careless student's way of solving the problem is also reliable. But, again, it seems incorrect to attribute knowledge to her. Why not? Well, she employed the right method on a whim, such that she could have very easily employed another, incorrect method.

So, how is this problem to be resolved? Is there a way to restrict the set of reliable belief-forming processes to those that are intuitively knowledge-conducive; i.e., to those processes that are responsive to the world, and which are neither strange nor fleeting? A more detailed analysis of what goes on in the cases just described should be illuminating.

1.3) The Ability Intuition and Virtue Reliabilism

1.3.1) The Ability Intuition

Let us focus on Hercules first. Hercules' way of forming his beliefs is perfectly reliable as his beliefs will systematically come out true. But we cannot attribute knowledge to him. The problem, as we noted before, is the direction of fit between his beliefs and the facts. In cases of knowledge, we want our beliefs to be true because they respond to the facts, and not because the facts comply with our beliefs; when one knows, one's true beliefs are about how the world is, not the other way around. In Hercules' case, however, his beliefs are not true because they are formed in a way that detects the facts. Instead, he first forms his beliefs in an arbitrary way—he makes no efforts to ensure that they will come out true—and then Zeus takes over so that the facts will comply with Hercules' beliefs. But this is not knowledge; this is the 'luck of the gods'. So, if, one day, Zeus had a fight with Hera, Hercules' beliefs would cease coming out true. But notice that if Hercules used his cognitive abilities—say, by taking a look at the sky—to form his weather beliefs, then he would not run into any such problems. If he formed his beliefs not in an arbitrary way, but on the basis of his cognitive abilities, he would not need Zeus to tweak the world so that his beliefs could systematically turn out true. If one's beliefs are the product of one's cognitive abilities they will be true because they are sensitive to the facts. In other words, the direction of fit will be correct.

So, it may be proposed that the way to restrict the reliable belief-forming processes to those that get the direction of fit correctly—such that they can be knowledge-conducive—is to identify them with one's cognitive abilities, or, in other words, with one's reliable cognitive processes. But can all reliable cognitive processes count as cognitive abilities? Not really, because, as the serendipitous brain lesion and careless student examples demonstrate, there are certain intuitions—which, as it will soon become apparent are closely connected to the issue of subjective justification—that disallow this move. Let me say more about them.

First, the serendipitous brain lesion demonstrates that there might be reliable cognitive belief-forming processes that we wouldn't like to claim they constitute cognitive abilities that can produce knowledge, because they are strange. More precisely, the intuition here is that for a cognitive process to count as a cognitive ability it must not be strange, in the sense that it must fit well with the rest of the agent's doxastic cognitive system. The reason is that if the cognitive process is strange, then, in light of the rest of his doxastic cognitive system, the agent will reject both the process and its deliverances, despite that it is in fact reliable—from the agent's point of view, it isn't. So, in order for a cognitive process to count as a cognitive ability such that it can be knowledge-conducive it must not be inconsistent with the rest of the agent's beliefs, and his methods for producing them. In other words it must be such that it can become part of, or, integrated within, the rest of the agent's doxastic cognitive system. But, clearly, this is not the case with the serendipitous brain lesion. The process is a cognitive malfunction, and, even more crucially, its output is so bizarre that no epistemic agent could accept as true. In other words, the serendipitous brain lesion cannot count as knowledge-conducive because it is so strange that it cannot be part of the rest of the agent's doxastic cognitive system, and so cannot count as a cognitive ability.

Second, the careless student's method of forming her true belief demonstrates that even a reliable cognitive process that is normal enough to become part of the rest of her doxastic cognitive system cannot yet count as a cognitive ability that can produce knowledge. The reason is that her reliable cognitive process of forming her true belief is a fleeting one; it is not a habit or a disposition of hers. In other words, she is in no sense aware that this is the way to solve the target problem; the reliability of the process is accidental from her point of view. Accordingly, given the same circumstances, the careless student could have so easily picked another, incorrect cognitive process for forming her beliefs, thus, ending up with a falsehood. If, instead, the student had habitually invoked the correct algorithm when the problems called for it, then we would indeed be inclined to claim that she can gain knowledge on its basis. The reason for this is that if a cognitive process is a disposition or a habit of the agent, then the agent will be able to become aware of the circumstances in which it can be unreliable. Otherwise, it seems arbitrary that the agent employed it in an appropriate, but isolated case, and so cannot gain knowledge on its basis. In other words, a reliable cognitive process that is normal—such that it can, in principle, become part of the rest of one's cognitive system—won't count as a cognitive ability, unless it is also a disposition or a habit of the agent. Why is this so? Well, the intuition is that abilities, in general, are habits or dispositions possessed by agents. But apart from such intuitions, we have also noted that in order for a reliable cognitive process to count as a cognitive ability it must be such that it can become part of (or, integrated within) the rest of the agent's doxastic cognitive system. One requirement for this, we have noted, is that the process not be strange such that it is not inconsistent with the rest of the agent's cognitive system. What is further required, however, is that it also be coherent with it. In other words, the agent must be able to become aware that it is unreliable in certain circumstances, so that she will be able to non-accidentally endorse it in the rest of the circumstances. And if the cognitive process is a disposition or a habit of the agent, then she can become aware of this.

So, we see that in cases of knowledge, we want one's way of forming one's beliefs to be responsive to the facts. Accordingly, we claimed that only reliable cognitive processes can be knowledge-conducive. But not all reliable cognitive processes will do; they also need to be normal dispositions, or habits of the agent so that they can become part of (or, integrated in) the rest of his doxastic cognitive system, such that they can count as cognitive abilities of the agent.

Greco (2004, 111), in a similar vain, claims that when we attribute knowledge to someone "we imply that the person is responsible for believing the truth", because believing the truth is the product of his cognitive abilities. Put another way, "to say that someone knows is to say that his believing the truth can be credited to him. It is to say that this person got things right due to his own abilities, efforts and actions, rather than due to dumb luck, or blind chance, or something else" (*Ibid.*).

Noticeably, the general idea, which all the above considerations are alluding to is that for a reliable process to be knowledge-conducive it must be a cognitive ability. This idea has also come to be known in the literature as the *ability intuition* on knowledge: *Knowledge must be the product of cognitive abilities*.[2] Moreover, the upshot of the above considerations is that for a process to be able to qualify as a cognitive ability, such that it can deliver knowledge, it must be a normal dispositional or habitual cognitive process. As we shall now see, it is exactly this understanding of the ability intuition on knowledge that when combined with process reliabilism gives rise to virtue reliabilism.

1.3.2) Virtue Reliabilism

So, in order to introduce virtue reliabilism, let us follow Greco who has proposed that not all reliable belief-forming processes are knowledge-conducive; rather "it is those processes that have their bases in the stable and successful dispositions of the believer that are relevant for knowledge and justification" (1999, 287). In other words, an epistemic agent *S* will be objectively justified in believing *p* just in case his true believing results from one of his dispositional reliable cognitive processes.

As Greco further notes (1999, 285), however, "it is not enough that one's belief is formed in a way that is objectively reliable; one's belief must be formed in a way that is subjectively appropriate as well". In other words,

[2] The idea that knowledge must be grounded in cognitive abilities can be traced back to the writings of Sosa (1988; 1993) and Plantinga (1993). For more recent approaches to the idea see Greco (1999; 2004; 2007) and Pritchard (2009; *forthcoming*; 2010*a*; 2010*b*).

Greco concedes that internalists are correct with respect to their intuition that one must be somehow sensitive to the reliability of the evidence one offers in favor of one's beliefs. Nevertheless, in order to remain fast to externalism such that he will also avoid Hume's skepticism, Greco suggests that subjective justification must be accommodated in a way that does not involve knowledge of, or even beliefs about reliability (see fn. 1). Accordingly, he proposes (1999, 289) that "a belief p is subjectively justified for a person *S* (in the sense relevant for having knowledge) if and only if *S*'s believing *p* is grounded in the cognitive dispositions that *S* manifests when *S* is thinking conscientiously" (i.e., when *S* is motivated to believe what is true). In this way the agent will employ his reliable cognitive processes in circumstances that have not been problematic in the past (i.e., those which are objectively reliable) and he will be able to do so without even having *any* beliefs about their reliability.[3]

In addition, Greco notes that the dispositions/habits that a person manifests when she is thinking conscientiously are the stable properties of her *cognitive character* (Greco 1999, 290). So, "a belief *p* has a positive epistemic status for a person *S* just in case *S*'s believing *p* results from the stable and reliable dispositions that make up *S*'s cognitive character" (Greco 1999, 287-8). In this way, Greco does away both with strange and fleeting processes. Strange processes cannot be part of the agent's cognitive character because they are not the kind of processes that a conscientious agent would employ. Fleeting processes are also excluded: First, because they are not dispositions or habits—so, they cannot really count as character traits. And, second,

[3] The fact that people manifest highly specific, finely tuned dispositions to form their beliefs in certain ways but not in others amounts to an implicit awareness of the reliability of those dispositions.

> For example suppose that it seems visually to a person that a cat is sleeping on the couch, and on this basis she believes that there is a sleeping cat on the couch. Suppose also that this belief manifests a disposition that the person has, to trust this sort of experience under these sorts of conditions, when motivated to believe the truth. Now, suppose that much less clearly, it seems visually to the person that a mouse has run across the floor. Not being disposed to trust this kind of fleeting experience, the person refrains from believing until further evidence comes in. The fact that the person, properly motivated, is disposed to trust one kind of experience but not the other, constitutes sensitivity on her part that the former is reliable. There is a clear sense in which she takes the former experience to be adequate to her goal of believing the truth, and takes the latter experience not to be. And this is so even if she has no beliefs about her goals, her reliability, or her experience (Greco 1999, 290) .

A similar argument can be found in (Sosa 1993, 60-63).

because it is only dispositions or habits that one can become aware they are unreliable in certain circumstances, and, so—without relying on any beliefs about their reliability—use them conscientiously in the rest of the circumstances.[4]

But, what might be part of one's cognitive character? On this view, one's cognitive character consists of one's cognitive faculties of the brain/central nervous system (CNS) including, of course, one's natural perceptual cognitive faculties, one's memories and the overall doxastic system. In addition, however, it can also consist of acquired habits of thought, "acquired skills of perception and acquired methods of inquiry, including those involving highly specialized training or even advanced technology" (1999, 287).

Virtue reliabilism is, therefore, a refinement of process reliabilism in that it accommodates the ability intuition on knowledge. In order for a belief to be both subjectively and objectively justified it's not enough that it is the product of a reliable belief-forming process; it must be the outcome of one of the agent's stable and successful *cognitive* belief-forming processes that make up his/her cognitive character. Accordingly, virtue reliabilism is usually formulated as follows:[5]

> *Virtue Reliabilism*
>
> *S* knows that *p* if and only if *S*'s reliable cognitive character is the most important necessary part of the total set of causal factors that give rise to *S*'s believing the truth regarding *p*.

As far as the ability intuition on knowledge is concerned, the thinking behind virtue reliabilism is this: if *S*'s true belief that *p* is the product of some

[4] So, given the discussion of the previous subsection, one can only employ a conscientious attitude towards the processes, which can, in principle, become part of (i.e., integrated in) one's cognitive character. In order, however, for one to indeed be conscientious towards them they must have actually been integrated within one's cognitive character. And this, as we shall see in chapter 3, further requires that the relevant processes densely interact with the rest of the agent's cognitive abilities. In a similar vein, Greco (2010, 152) writes: "in general, it would seem, cognitive integration is a function of cooperation and interaction, or cooperative interaction, with other aspects of the cognitive system".

[5] Notice that Greco calls his view 'agent reliabilism'. I have here preferred this alternative name for two reasons. First, in the article that we have so far been closely following, Greco does not clearly present his view as a complete account of knowledge. Second, I wanted to make explicit that this approach falls under the broader trend of virtue epistemology, "since the stable and successful dispositions of a person are appropriately understood as virtues" (Greco 1999, 287).

cognitive ability, then we may conclude that *S's* cognitive character figures most importantly in the causal explanation of how *S* came to believe the truth regarding *p*, and thus that *S* knows that *p*.

Notice, moreover, that the reason why virtue epistemologists are inclined towards such a strong virtue-theoretic account of knowledge is their attempt to do away with the knowledge-undermining epistemic luck involved in Gettier cases.[6] As Gettier demonstrated, one's justified belief may turn out to be true without thereby counting as an instance of knowledge. In the typical scenario, one's belief, which is the product of faulty reasoning, *just happens* to be true for reasons that are extraneous to one's justification. Or again, one may come to believe the truth on the basis of a lucky guess. Contrast this with cases of success through ability. "There is a sense of "luck" on which lucky success is precisely opposed to success through virtue or ability" (Greco 2007, 58). When one's success is the product of one's ability then clearly one's success cannot have been a lucky one. Accordingly, virtue epistemologists hold that when one knows, one's intellectual success is the product of cognitive ability.[7] In other words, they claim, the cognitive success must be primarily creditable to one's cognitive character. (Alternatively: one's cognitive character must be the most salient factor in the causal explanation of how one acquired one's true belief).[8]

So, let us now direct our attention to several thought experiments that will help us properly evaluate virtue reliabilism. First, consider Hercules. Obviously, Hercules' beliefs are not the result of his cognitive abilities, so his

[6] See (Gettier 1963).

[7] Notice, here, that the claim is that the cognitive success must be the *product* of cognitive ability. It is not the weaker claim that cognitive ability must have been involved in the acquisition of one's true belief, since this can be satisfied far too easily in ways that do not exclude luck.

[8] Obviously, this is a 'causal explanatory' reading of virtue reliabilism. There is a, however, one more standard understanding of virtue reliabilism available in the literature. Roughly, according to this second understanding, the view is formulated by demanding that one's cognitive success be *because of* one's cognitive ability, where the 'because of' relation between true belief and ability is not understood in purely causal terms, but on the basis of the 'dispositions manifestation' model of explanation. See, for example, (Sosa 2007: ch. 5). Although these two understandings of virtue reliabilism will, in most cases, produce the same results, notice that dispositions are second-order properties that derive from further properties of their bearers. For example, the fragility of a vase depends on the molecular structure of the vase. Accordingly, if we tried to understand why the vase broke, the 'dispositions manifestation' model of explanation would stop at the fragility of the vase, whereas a 'causal explanatory' model would penetrate further, by referring to the molecular structure of the vase. It could be the case, therefore, that these two formulations of virtue reliabilism may, sometimes, produce different results.

cognitive character has nothing to do with his believing the truth regarding the temperature of the room. On the contrary, it is Zeus' intervention that is the most salient factor in the causal explanation of how Hercules believes the truth. Hence, virtue reliabilism rules correct that Hercules lacks knowledge.

Or, consider Roddy:

> Roddy (Pritchard 2009, 11)[9]
>
> Roddy is a farmer. One day he is looking into a field near-by and clearly sees something that looks just like a sheep. Consequently he forms a belief that there is a sheep in the field. Moreover, this belief is true in that there is a sheep in the field in question. However, what Roddy is looking at is not a sheep, but rather a big hairy dog that looks just like a sheep and which is obscuring from view the sheep standing just behind.

Roddy's cognitive success cannot be attributed to his cognitive character. The reason is that Roddy's actions (taking a look at a sheep-shaped dog) are not the most important part in the correct causal explanation of how he believes the truth. Instead, what is salient for explaining his success is the abnormal presence of a great amount of epistemic luck (i.e., there being a real sheep behind the sheep-shaped dog). According to virtue reliabilism, then, and in line with our intuitions, Roddy cannot gain knowledge in this way. Before moving on, however, let me note a subtle detail, which is of great importance to the formulation of virtue reliabilism. The demand is that one's belief be *true because of/in virtue of* one's cognitive character. Virtue reliabilism rejects the weaker claim that one must believe on the basis of one's cognitive character *and* that one's belief happens to be true. Such a weaker claim could not accommodate the Roddy case.

So far so good for virtue reliabilism. The following counterexample, however, shows the view to be too weak an account of knowledge.

> Barney (Pritchard 2009, 12)[10]
>
> Barney is driving through the country and happens to look out of the window into a field. In doing so, he gets to have a good look at a barn-shaped object, whereupon he forms the belief that there is a barn in the field. This belief is true, since what he is looking at is really a barn. Unbeknownst to Barney, however, he is presently in 'barn-façade country' where every object

[9] The Roddy case is described in Chisholm (1977, 105).

[10] The Barney case is described in Goldman (1976) and credited to Ginet.

> that looks like a barn is a convincing fake. Had Barney looked at one of the fake barns, then he would not have noticed the difference. Quite by chance, however, Barney just happened to look at the one real barn in the vicinity.

Barney comes to truly believe that he is looking at a real barn by employing his cognitive abilities; he is looking directly at the barn. Therefore, we now have a case where despite the fact that Barney's true belief is solely formed on the basis of his cognitive ability, his cognitive success cannot be called knowledge given that his belief could have so easily been false (Barney is in a barn-façade environment). Put another way, the objection is that we now have cognitive success that does not amount to knowledge even though it is solely formed on the basis of Barney's reliable cognitive character.

Consequently, it has been argued that virtue reliabilism cannot ultimately do all the work that is expected to do. Despite virtue epistemologists' initial expectations, the ability intuition on knowledge seems unable to *fully* accommodate the equally important intuition that knowledge must not be due to luck, *viz.*, the anti-luck intuition on knowledge. Instead, it seems that in order to *fully* deal with knowledge-undermining luck, virtue reliabilists must also incorporate a separate anti-luck condition into their theory of knowledge.[11] The upshot appears to be that these two intuitions about knowledge "impose independent epistemic demands on our theory of knowledge" (Pritchard *forthcoming*).

Accordingly, friends of anti-luck epistemology claim that any adequate theory of knowledge must explicitly have as a central component an anti-luck epistemic condition such as the safety or the sensitivity principle.[12] In contrast

[11] Notice, here, a subtle difference: In regular Gettier cases, the knowledge-undermining luck is very direct, in that the luck concerns the relationship between the belief and the fact. One's belief is erroneously formed but there is a lucky fact that renders it true. Luck intervenes between the belief and the fact. On the contrary, in cases like the one Barney is in, the knowledge-undermining luck is quite indirect; indeed, it is specifically *environmental*. Barney does look at a real barn and believes that there is a real barn in front of him. There is nothing wrong with the way he forms his belief. Given, however, the environmental conditions he cannot acquire knowledge in this way. Luck interferes with the environment in such a way that even a well-formed belief will be a lucky one if true. So, it seems that while the ability condition on knowledge can deal with the normal (non-environmental) knowledge-undermining *epistemic* luck involved in Gettier cases, it may not be able to deal specifically with *environmental* knowledge-undermining luck. See also (Kallestrup & Pritchard (*forthcoming*). I have elsewhere argued that a virtue reliabilistic condition on knowledge can deal with all sorts of knowledge-undermining luck (Palermos *forthcoming*).

[12] The *sensitivity principle* is usually formulated as follows: If *S* knows that *p*, then *S*'s true belief that *p*, is such that, had *p* been false, *S* would not have believed *p*. The classic defenses of the sensitivity principle can be found in Dretske (1970) and Nozick (1981). The *safety*

to the ability condition on knowledge, which arguably addresses the problem posed by knowledge-undermining luck only indirectly, these modal conditions on knowledge are primarily targeted to capture the anti-luck requirement, as they are explicitly concerned with the responsiveness of one's belief to relevant counterfactual circumstances (such as the scenario in which Barney looks at a barn-façade instead of a real barn). Arguably, then, for virtue reliabilism to be a fully adequate account of knowledge it may have to be supplemented by a specific anti-luck condition on knowledge such as safety or sensitivity.[13]

On the face of the Barney counterexample, therefore, virtue reliabilism appears to be an insufficient condition on knowledge. The following case, however, demonstrates that it is too strong as well.

> Jenny (Pritchard 2009, 68) [14]
>
> Jenny gets off the train in an unfamiliar city and asks the first person that she meets for directions. The person that she asks is indeed knowledgeable about the area, and gives her directions. Jenny believes what she is told and goes on her way to her intended destination.

Now, unless we want to deny a great amount of knowledge that we suppose we have, we must admit that Jenny gains knowledge in this way. Given the way Jenny gains knowledge, however, her cognitive character, it seems, has nothing to do with the truth status of her belief. Instead, it is the informant's cognitive character that is the most salient (and maybe the only) factor in the causal explanation of why Jenny believes the truth. So, according to virtue

principle is usually understood thusly: if *S* knows that *p*, then *S'*s true belief that *p,* is such that *S's* belief that *p* could not have easily been false. For recent defenses of the safety principle see Sosa (1999; 2000) and Pritchard (2002; 2008). For a very good discussion concerning the relation between the ability and the anti-luck intuition on knowledge see Pritchard (*forthcoming*).

[13] Consider for example *Anti-Luck Virtue Epistemology: S knows that p if and only if S's safe belief that p is the product of her relevant cognitive abilities (such that her safe cognitive success is to a significant degree creditable to her cognitive agency)* (Pritchard *forthcoming*, 20). Again, in (Pritchard 2010*a*, 76) we can read: " knowledge is safe belief that arises out of the reliable cognitive traits that make up one's cognitive character, such that one's cognitive success is to a significant degree creditable to one's cognitive character".

[14] The Jenny case is adapted from Lackey's 'Morris case' (2007, 352). As the thought experiment is here laid out, nothing is really changed, apart from the hero's name. As we shall see in the next section, however, Pritchard favors a slightly different—and admittedly more intuitive—understanding of the case.

reliabilism Jenny lacks knowledge that she in fact possesses.[15]

This counterexample shows virtue reliabilism to be too strong as it stands, and that it should therefore be somehow weakened. Accordingly, it is now time to move on to the consideration of Pritchard's recent attempt to capture the ability intuition by putting forward a necessary condition on knowledge, namely COGA$_{\text{weak}}$ (2010*b*).[16]

1.4) COGA$_{\text{weak}}$

1.4.1) COGA$_{\text{weak}}$

Before quoting COGA$_{\text{weak}}$, however, we should first take a look at which considerations have led to it by investigating the Jenny case in some more detail. First, as Pritchard explains, to say that Jenny gains knowledge, we must read the example in such a way that Jenny is in an epistemically friendly environment—i.e., the city that Jenny visits had better not be renowned for its dishonest informants. Were that the case, we would not credit Jenny with knowledge. Second, notice that we presuppose some natural inclinations about Jenny's cognitive character. We expect that Jenny can distinguish between potentially reliable and clearly unreliable informants; we do not expect that Jenny would be happy to ask just anybody. For example, we anticipate that she would not ask someone who clearly looked like a tourist (i.e., an unreliable informant). "Had the first person she met been obviously mad, or a stereotypical tourist, for example, then we would expect her to move on to the next prospective informant down the street" (Pritchard

[15] Traditionally, accounts of testimonial knowledge are divided in two main trends. The first one is called reductionism and it is the view that a hearer is justified in believing a speaker's testimony if and only if she has non-testimonial positive reasons in favor of the speaker's reports, such that her justification for accepting them is reducible to basic sources of knowledge such as sense perception, memory and inductive inference. Some of the proponents of reductionism are considered to be Hume (1977), Faulkner (2000), Fricker (1994). The second trend is called non-reductionism and it is the view that one is by default justified in believing one's testimony unless one has negative reasons for doing so. In this view, testimonial justification cannot be reduced to more basic sources of knowledge. Typical proponents of non-reductionism on testimonial knowledge are thought to be, amongst others, Reid (1983), Burge (1993), Weiner (2003) and Audi (1998). Recently, however, Jennifer Lackey (2008) has put forward a dualist account of testimonial knowledge, which accommodates both of the aforementioned views.

Virtue reliabilism appears to accord only with reductionism on testimonial knowledge, due to its strong demand that the cognitive success should be primarily creditable to the hearer's cognitive character.

[16] Pritchard argues for this condition in a number of places, though under different names and slightly different formulations. See (Pritchard *forthcoming*, 20), (Pritchard 2009, 74) and (Pritchard 2010*a*, 76).

forthcoming, 18). Moreover, we expect that she is able to distinguish between potentially reliable and clearly unreliable information and thereby that she would not believe whatever she was told, had it been obviously false (for instance to go past the city hall whereas, in fact, she is in a village). "Furthermore, if the manner in which the informant passed on the directions was clearly questionable—if the informant was vague, shifty, hostile, and evasive, say—then we would expect our hero to exercise due caution" (Pritchard *forthcoming,* 18). In other words, had Jenny not been responsive to these epistemologically relevant factors we would not have normally attributed knowledge to her. We, therefore, see that it is not that Jenny's cognitive character has nothing to do with her believing the truth; it is just that the informant's role is more important. It is upon these considerations that Pritchard proposes COGA$_{\text{weak}}$.[17]

> *COGA$_{weak}$*
>
> If *S* knows that *p*, then *S*'s true belief that *p* is the product of a reliable belief-forming process, which is appropriately integrated within *S*'s cognitive character such that her cognitive success is to a significant degree creditable to her cognitive agency. (Pritchard 2010*b*, 136-7)

Obviously COGA$_{\text{weak}}$ can easily handle the Jenny case; although the cognitive success is not *primarily* creditable to Jenny's cognitive character—but rather the stranger who delivers the reliable information—Jenny, in so being responsive to the epistemologically relevant factors, has the right sort of abilities and employs them in the right sort of way so as to accept the stranger's information, such that believing the truth is *significantly* creditable to her cognitive agency. According to COGA$_{\text{weak}}$, then, Jenny can gain knowledge in this way.

What is of import here is to notice the lenient demands of COGA$_{\text{weak}}$ regarding the creditability of the cognitive success to one's self. In contrast to virtue reliabilism where believing the truth must be primarily creditable to one's cognitive character and thereby to one's self, COGA$_{\text{weak}}$ loosens the required dependence of the cognitive success on one's *cognitive agency* thereby

[17] Notice that Pritchard avoids making COGA$_{\text{weak}}$ a complete account of knowledge because, having in mind counterexamples such as the Barney case, he holds that for an adequate account of knowledge, COGA$_{\text{weak}}$ must be supplemented by an anti-luck condition on knowledge such as the safety principle.

allowing significant part of the credit to be attributable to other factors as well. Therefore, according to COGA$_{weak}$, even though the most salient factor that explains Jenny's cognitive success is the informant's contribution, Jenny's cognitive abilities render her cognitive agency significantly creditworthy, thereby allowing her to gain knowledge in this way.[18]

Put another way, since Jenny employs the relevant belief-forming processes in order to rationally accept the speaker's report, believing the truth is *significantly*—though not primarily—creditable to her cognitive agency and therefore, just as COGA$_{weak}$ allows, Jenny can gain knowledge in this way. Crucially, however, the rest of the credit should be, in accordance to our intuitions, attributed to the speaker's cognitive agency for delivering reliable information. Therefore, COGA$_{weak}$ has the means to explain the dual sources of justification in cases of testimonial knowledge by attributing credit to both parties of the said exchange.[19]

Furthermore, COGA$_{weak}$ rules correct with respect to all the previous thought experiments that we have so far encountered. Hercules lacks knowledge, because believing the truth is in no degree creditable to his cognitive agency but, instead, to Zeus. Moreover, neither the agent with the serendipitous brain lesion, nor the careless math student possesses knowledge. The reason is that their reliable belief-forming processes cannot become, or do not yet count, respectively, as parts of (i.e., they cannot be, or they have not yet been, appropriately integrated within) their cognitive characters. And, finally, the only worth-mentioning factor in the causal explanation of how Roddy believes the truth is the abnormal presence of luck that there happens to be a real sheep behind the sheep-shaped dog. Consequently, believing the truth cannot be significantly credited to his cognitive agency.[20]

[18] Notice, then, that the cognitive success being *primarily* creditable to one's cognitive agency and it being the product of one's cognitive abilities is not exactly the same thing. That is, one's cognitive success can be the product of one's cognitive abilities even if it is not *primarily* creditable to one's cognitive agency.

[19] For a more detailed analysis of COGA$_{weak}$ along the lines of dualism in the epistemology of testimony, as put forward by Lackey (2008), see (Palermos 2011). See also section 4.2.

[20] The Barney case should not be a problem for COGA$_{weak}$ which is only a necessary condition on knowledge; even if Barney's cognitive success can be significantly credited to his cognitive agency, he may nevertheless lack knowledge because he fails to satisfy some supplementary condition on knowledge such as the safety principle.

1.4.2) Cognitive Agency, Epistemic Artifacts, and the Ability Intuition

Having seen how $\text{COGA}_{\text{weak}}$ is able to overcome the problems that its predecessors face, I now want to discuss a both presentational and explanatory advantage of the view, namely the explicit recognition of the central role that one's cognitive agency plays in the knowledge-acquiring process. Before that, however, notice that what virtue reliabilism and $\text{COGA}_{\text{weak}}$ have in common is their attempt to understand knowledge, or at least a necessary aspect of it—recall that $\text{COGA}_{\text{weak}}$ is not supposed to be a sufficient condition on knowledge—in terms of credit attributions. This is so because in trying to accommodate the ability intuition on knowledge, both views share another common idea: credit is usually attributable in cases of success through ability.[21]

Greco (2004) claims that credit attribution involves ascribing action and ascribing action involves causal citation of the subject who performed the relevant action. Accordingly, any adequate account of intellectual credit attribution should—or so I suggest—not only refer to the person's cognitive character, but, also, to the person's cognitive agency (which *acted* on the person's cognitive character for getting things right). If this is right, then the formulation of $\text{COGA}_{\text{weak}}$ is clearly advantageous in comparison to virtue reliabilism, since the former condition, in contrast to the latter, explicitly demands that one's true belief be creditable to one's cognitive agency and not merely to one's cognitive character. It may be objected, however, that this should not be more than a merely presentational concern.

This worry brings me to the second point I want to make concerning the explanatory merits of $\text{COGA}_{\text{weak}}$. Recall that in addition to one's organismic cognitive abilities of the brain/CNS, a person's cognitive character

[21] A subtle difference between the two proposals, however, is that while Greco presents knowledge as true belief which is 'of credit', Pritchard insists on thinking about knowledge merely as 'creditable' true belief. These two notions are not the same. "For example, one's cognitive success could be creditable to one's cognitive agency without being at all of credit to one (perhaps the cognitive success is the result of an inquiry that one ought not to be pursuing, because, say, there are epistemically more desirable inquiries that one should be focusing instead" (Pritchard *forthcoming*, en. 26). While this distinction is not important to the present discussion, it is of great significance with respect to the debate on the value of knowledge. If, as Greco claims, knowledge is true belief, which is 'of credit', this is because knowledge is an achievement. Since achievements are finally valuable, knowledge turns out to be finally valuable, as well. Considering, however, cases such as the one mentioned above, or mundane instances of knowledge such as perceptual beliefs, Pritchard claims that knowledge is not always an achievement and so not finally valuable either. For further discussion on this issue, see (Pritchard 2010*a*, §2.4).

may also consist of "acquired skills of perception and acquired methods of inquiry, including those involving highly specialized training or even advanced technology" (Greco 1999, 287). The reason for this move is that virtue reliabilism needs to also account for advanced cases of knowledge in which one's believing the truth is the product of the operation of epistemic artifacts such as telescopes, microscopes, and so on. In the traditional conception, however, cognition takes place strictly within an agent's head and so artifacts cannot be parts of one's *cognitive* character.

One way to sidestep this problem for virtue reliabilism could be to claim that, in such cases, it is merely the agent's training and skill of using the artifact, as mirrored in the agent's neural/bodily architecture, that is the most salient factor in the causal explanation of acquiring the cognitive success. Notice, however, that when an agent employs an epistemic tool, his true belief arises as the product of the *interaction* between his internal processes and the artifact. What this means is that the agent's cognitive process that allows him to detect the truth is not merely 'aided' or 'assisted' by the artifact, but, is, instead, constituted by it, as it arises out of his ongoing engagement with the artifact. It thereby appears that it will be impossible to disentangle the agent's training and skill of using the artifact from his actual engagement with it, in a causal explanation of how the agent acquired his true belief.[22]

But even if such decomposition were possible, notice in addition that the part of the process that allows the cognitive agent to detect the truth, or in other words to be sensitive to the facts, is the external component. To make this point clear, consider, on one hand, an untrained agent in possession of a properly working artifact. In that case it is obvious that even though the agent will initially be unable to form any (true or false) beliefs, eventually—provided that he gains sufficient experience—not only will he form beliefs, but he will also reliably enjoy cognitive success. On the other hand, think about a well-trained agent, but in possession of a faulty artifact. In this case, despite the agent's excellent internal skills, it is evident that he would be unable to reach any (non-lucky) true beliefs, no matter how much he tried. It, therefore, seems that, in such cases, the only significant factor that explains

[22] Provisionally this point may seem ambiguous but it will hopefully become clearer in the discussion of chapter 2 where I explore the consequences of the phenomenon of continuous reciprocal causation between the organismic agent and her epistemic artifact.

the agent's cognitive success is the epistemic artifact.[23] In other words, since the epistemic artifact is the only significant part of the *integrated* belief-forming process that produces one's *true* believing, the virtue reliabilist must account for it being part of one's cognitive system. Given, however, that cognition is normally supposed to take place within the agent's head, Greco has no principled way to show why such artifacts may count as proper parts of one's cognitive character. Therefore, there seems to be a worrying tension between the ability intuition on knowledge and such a broad understanding of the notion of one's cognitive character.

According to $\text{COGA}_{\text{weak}}$, however, so long as one forms true beliefs on the basis of some process in such a way that believing the truth can be significantly creditable to one's cognitive agency, then one may be said to know the target proposition. Since $\text{COGA}_{\text{weak}}$ is a formulation of the ability intuition on knowledge, what this means is that so long as one's true belief is significantly creditable to one's cognitive agency, then the process by which one came to form one's belief can be said to have been appropriately integrated within one's cognitive character, and thereby count as a *bona fide* cognitive ability.[24] Therefore, in contrast to virtue reliabilism, $\text{COGA}_{\text{weak}}$ with its appeal to one's cognitive agency offers a principled way to account for the acquisition of knowledge on the basis of epistemic artifacts, while retaining the ability intuition on knowledge.

Consequently, we see that far from being merely a presentational point of consideration, the inclusion of the notion of one's cognitive agency in the formulation of $\text{COGA}_{\text{weak}}$ does a great deal of explanatory work. That is, apart from directing our knowledge attributions to the primarily responsible element of the knowledge-forming process—and not just to the subject's cognitive character—$\text{COGA}_{\text{weak}}$ can also create some important logical space for explaining how cognitive agents may come to gain knowledge on the

[23] This is not to claim that one's internal processes are not a significant factor in how one forms one's beliefs. The only significant factor, however, in how one *believes the truth of the matter* is the artifact.

[24] Pritchard makes this point in (2010*b*, 137; en. 7). Although this might generally be a good way to start judging whether a process can count as a *bona fide* cognitive ability, notice that Pritchard uses the terms 'cognitive agency' and 'cognitive character' interchangeably thereby running the risk of rendering his criterion circular or at least unsafe. As I argue in section 3.3, where I discuss several thought experiments, Pritchard has indeed been led astray by his criterion with respect to one version of the Temp case. Therefore, we are in need of an alternative way to judge whether an external process has been *appropriately* integrated within one's cognitive character. I offer such an alternative, which is nevertheless in the spirit of Pritchard's suggestion, in section 3.2.

basis of processes which lie outside their heads but which, nevertheless, may be proper parts of their cognitive characters.

1.5) Conclusion

In this chapter we focused on the motivation for a theory of knowledge that accommodates the recognition of the contingent reliability of our evidence. Process reliabilism initially seemed to fit the bill, but we soon realized that a plausible account of knowledge must also ensure the correct direction of fit between one's beliefs and the facts, as well as deny that strange and fleeting processes can give rise to knowledge. Accordingly, process reliabilism, we claimed, must be enriched with the ability intuition on knowledge (i.e., knowledge must be the product of cognitive ability), thus giving rise to virtue reliabilism. The latter proposal, however, turned out to be a both too strong and weak account of knowledge on the basis of several counterexamples. This, eventually, led us to the embracement of COGA$_{weak}$.

Furthermore, apart from accommodating our intuitions with respect to the thought experiments that trouble virtue reliabilism, COGA$_{weak}$ can also create some important logical space for explaining how we may come to acquire knowledge on the basis of epistemic artifacts, which lie outside our heads, but which may nevertheless count as parts of our cognitive characters. Considering, however, artifacts as part's of one's cognitive character such that their employment can count as a *bona fide* cognitive ability is a rather radical claim, which had better not been left unsupported.

Accordingly, there remains to see whether there are available any independent resources for properly conceptualizing the use of artifacts in a way that is continuous to our understanding of organismic cognitive abilities.[25] And should there be such an account, what does it take for those artifacts to be appropriately integrated within (i.e., be proper parts of) one's overall cognitive character, other than our intuitions on the degree of creditability of one's cognitive success to one's cognitive agency? Moreover, the notions of one's 'cognitive character', 'cognitive agency' and 'reliable belief-forming processes' (i.e., cognitive abilities), which are all central to the

[25] That is, even though we might be epistemologically motivated to accept the employment of artifacts as *bona fide* cognitive abilities, such a claim would be very weak in the absence of any metaphysical support.

formulation of COGA$_{weak}$, have only been vaguely discussed.

The following chapter will try to set the framework that will allow us to disambiguate all the above points. For this reason we will concentrate on contemporary philosophy of mind and, in particular, on the Hypothesis of Extended Cognition (HEC), as it has been put forward by Clark and Chalmers (1998).

CHAPTER 2

The Hypothesis of Extended Cognition

2.1) Introduction

Most probably, the only way to argue for the inclusion of artifacts within one's cognitive character is through the consideration of the hypothesis of the extended cognition (HEC). According to HEC, "the actual local operations that realize certain forms of human cognizing include inextricable tangles of feedback, feedforward and feed-around loops: loops that promiscuously criss-cross the boundaries of brain, body and world" (Clark 2007, sec. 1).

HEC is a form of active externalism, which should be distinguished from meaning (or passive) externalism as presented in the writings of Putnam (1975) and Burge (1986). Active externalism, which is different from the aforementioned traditional forms of externalism in that it concerns the aspects of the environment that *drive* one's cognitive loops in an ongoing way has been defended by many philosophers and appears in the literature with as many names as the number of its proponents: the extended mind (Clark and Chalmers 1998), environmentalism (Rowlands 1999), locational externalism (Wilson 2000; 2004), cognitive integration (Menary 2007) and so on. However, for reasons of simplicity and theoretical affinity to the epistemological framework of virtue reliabilism—as it will be explored in the chapter to follow—I will here discuss active externalism only on the basis of the terminology and argumentative lines that appear in Clark and Chalmers' initial proposal (1998).

To start with, consider first the well-known arcade game Tetris. In order to decide where the falling pieces (two-dimensional geometric shapes) will best fit in changing 'sockets' at the bottom of the screen, the player has two options: either (a) perform in his imagination a mental rotation or (b) use an onscreen button that causes the falling piece to rotate. Imagine, however, a third case (c): in the future, the agent is equipped with a neural implant, which can perform the rotation operation as fast as the rotation button in case (b).

Let us now see what the relation between these cases is. Clark and

Chalmers claim that the third case of the neural implant seems to be on a par with case (a). Moreover, the second case of the rotation button manifests the same computational structure as case (c) despite the fact that it is distributed across the agent and the onscreen button.[26] Therefore, if one treats case (c) as cognitive—since it is on a par with case (a)—there is no principled way to deny that case (b), where the process of decision by rotation is distributed across the player and the computer, is cognitive as well. For if one points to the skin and skull boundary, she will beg the question as this is exactly what is at issue here.

With this thought experiment in mind, Clark and Chalmers propose (1998, 8) the following principle as a way to challenge our intuitions on the cognitive status of some process:

> Parity Principle
>
> If, as we confront some task, a part of the world functions as a process which, were it go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process.

The essence of the parity principle is this: whether a process should count as a cognitive one must not depend on spatial considerations (i.e., whether it lies beyond one's skin) but, instead, on whether it plays 'the right kind' of active role in driving one's overall cognitive loops.[27]

Moreover, according to HEC, when parts of the environment become properly coupled to an agent's brain, they can be considered as constitutive parts of the overall cognitive mechanism—*viz.*, cognition potentially extends to the world surrounding the agent. Put another way, "in these cases, the human organism is linked with an external entity in a two-way interaction creating a *coupled system* that can be seen as a cognitive system in its own right" (Clark & Chalmers 1998, 8). Clark has also termed this two-way interaction as 'continuous reciprocal causation' (CRC): "CRC occurs when

[26] That is, instead of being wholly realized within the player's head.

[27] Where 'the right kind' of role is to be determined by a commonsense functionalist understanding of the relevant cognitive task. In the Tetris case, for example, the cognitive task is the 'decision where to place the falling piece by rotating it', and the claim is that since the interaction between the agent, the onscreen button and the visual perception of the rotating piece plays the same functional role in deciding where to place the falling piece as 'mental'/imaginary rotation does, then the externally delegated version of the process should count as cognitive as well.

some system S is both continuously affecting and simultaneously being affected by activity in some other system O" (Clark 2008, 24). In such cases, according to Dynamical Systems Theory, in order to model the temporal evolution of the two systems S and O, one must further postulate a *coupled* system E consisting of both S and O. So the claim is that, in cases of extended cognition, one's internal cognitive capacities are combined (i.e., mutually interact) with some environmental element O to form an extended cognitive whole, E, whose behavioral competence will drop if one removes the external component, just as it would drop if one removed part of its brain.

Consider, for example, the use of pen and paper when trying to solve a complex, say, a three-digit multiplication problem such as 987 times 789. It is true that few, if any, of us can solve this problem by looking at or contemplating on it. We may only perform the multiplication by using pen and paper to externalize the very problem in symbols. Then, we can serially proceed to its solution by performing simpler multiplications starting with 9 times 7. In this way, the pen and paper compensate for our limited working memories allowing us to perform a task that is otherwise infeasible. If one should try to describe how a regular human mind may perform such a cognitive task then, apart from the states and properties of a typical human brain, one should also factor in both the normative aspects of the notational/representational system involved, and the properties and *ongoing* states of the mediums with which the manipulation of the representations was performed.[28]

Or, think about the role of language when writing a philosophy paper. According to Clark, language too is "an external epistemic artifact designed to complement, rather than recapitulate or transfigure, the basic processing profile we share with other animals" (1998, 169). As I write down this essay,

> I am continually creating, putting aside, and re-organizing chunks of text. I have a file, which contains all kinds of hints and fragments, stored up over a long period of time, which may be germane to the discussion. I have source texts and papers full of notes and annotations. As I (literally, physically) move these things about, interacting first with one, then another, making new notes, annotations and plans, so the intellectual shape of the chapter grows and solidifies. It is a shape which does not spring fully developed from inner cogitations. Instead, it is the product of a sustained and iterated sequence of

[28] For the importance of the normative aspects of the external representational systems in explaining cognition see (Menary 2007).

> interactions between my brain and a variety of external props (Clark 1998, 173).

The moral, Clark claims, is that public language and text play more than just a preserving-and-communicating-ideas role:

> Instead, these external resources make available concepts, strategies and learning trajectories which are simply not available to individual un-augmented brains. Much of the true power of language lies in its underappreciated capacity to re-shape the computational spaces which confront intelligent agents (Clark 1998, 173).[29] [30]

To return, now, to the discussion of the core tenets of HEC, let us take a look at a further example that Clark and Chalmers (1998) use in order to argue for the more provocative Extended Mind Thesis from the existence of extended mental states.[31] First, think about a normal case of a belief stored in biological memory. Inga learns about an interesting exhibition in MOMA. She thinks, recalls that the museum is on 53rd street and starts walking to the museum. Now take a look at Otto who suffers from Alzheimer's disease; as a consequence, Otto has to rely on information in the environment to help structure his life and, thus, carries a thick, well-organized notebook everywhere he goes; when he learns new information he writes it down, when he needs some old information he looks it up. Otto hears about the

[29] For a discussion on Clark's view regarding language see (Wheeler, 2004). For a straightforward criticism, see (Rupert, 2010).

[30] Moreover, if one allows for such an understanding of public language and text, then important conceptual space is created so as to lengthen the list of epistemic artifacts in interesting ways. For, as Robert Logan (2003, 275) claims, "speech, writing, math, science and computing form an evolutionary chain of languages. Each of these activities can be considered as a separate language because each allows us to think differently, create new ideas and develop new forms of expression. Another consideration is that each of these five forms of language possesses its own unique semantics and syntax and hence qualifies as a language in itself according to criteria set by classical linguistics". While much more remain to be said on this matter, concentrating on the case of scientific theories, it is interesting that philosophers of science such as Imre Lakatos write: "[scientists] *use our most successful theories as extensions of our senses*" (Lakatos 1970, 107, emphasis in the original). Given the appropriate theorizing, however, this may turn out to be more than just a metaphor. That is, it could be the case that scientific theories, like public language and text, are software epistemic artifacts that extend one's cognitive abilities beyond one's natural cognitive capacities. I will return to this claim in section 3.4.

[31] Given that cognitive processes are mental processes, HEC can also lead to the extended mind thesis. I here say, however, that the latter is more provocative, because its motivation in the literature relies on extended mental states, such as beliefs. The existence of extended mental states, however, is a claim that is admittedly more counterintuitive and much less easier to motivate than the claim that there are extended cognitive processes.

same exhibition and decides to go see it. He opens the notebook, finds the address of the museum and starts heading towards 53rd street.

Clark and Chalmers claim that Otto walked to 53rd street because he wanted to go to MOMA and believed that MOMA was on 53rd street. What is more, if one wants to say that Inga had her belief before she consulted her memory, then one could also claim that Otto believed that the museum was on 53rd street even before looking up the address in his notebook. This is because the two cases are functionally on a par;[32] "the notebook plays for Otto the same role that memory plays for Inga; the information in the notebook functions just like the information [stored in Inga's biological memory] constituting an ordinary non-occurent belief; it just happens that this information lies beyond the skin" (1998, 13).

Although the postulation of extended mental states is not necessary for making the case for HEC, thereby allowing us to bypass the admittedly long debate that Otto has generated, the discussion of this example is helpful as it has produced some very important intuitions on what is required for an external artifact to count as a putative part of one's overall cognitive economy. In particular, investigating the case in more detail, Clark (2010*a*) notes that the availability and portability of the resource of information might be crucial. Accordingly, he offers the following set of additional criteria to be met by non-biological candidates for inclusion into an individual's cognitive system (2010*a*, 46)[33]:

> 1) "That the resource be reliably available and typically invoked".
> 2)"That any information thus retrieved be more-or-less automatically endorsed. It should not usually be subject to critical scrutiny. […] It should be deemed about as trustworthy as something retrieved clearly from biological memory".
> 3)"That information contained in the resource should be easily accessible as and when required".

These criteria have also come to be known in the literature as the 'glue and trust' criteria and they are primarily meant to ensure the effect of 'transparent

[32] That is, they are 'functionally on a par' as explained in fn. 27.

[33] This paper was first published in *The Extended Mind,* (2010), Menary (ed.) Cambridge, Massachusetts, MIT press, but it has been available online since 2006. The 'glue and trust' criteria, however, had already made their appearance in (Clark and Chalmers 1998), although the phrasing was somewhat different.

equipment': "equipment (like the carpenter's hammer) with which we are so familiar and fluent that we do not think about it in use, but rather rely on it to mediate our encounters with a still-wider world" (Clark 2006, 106). Put another way, an external element is part of one's ongoing cognitive loops when it is not part of the problem space but is instead one of the mediums manipulated in order to complete the cognitive task at hand.

2.2) Objections

It is true that in its relatively short history HEC has faced an impressive number of objections. In this section, the discussion will focus on some specific objections which, even though have not been deemed particularly worrying, will help us familiarize with the view. This will allow us, in the next section, to address two distinctively important, yet interrelated objections—namely the 'coupling-constitution' fallacy and the 'cognitive bloat' worry—whose discussion, as it will gradually become apparent, is closely connected to the aim of delineating the hard core of HEC. Proceeding in this way will also help us to distinguish HEC from a superficially similar, yet fundamentally different view, that of the hypothesis of embedded cognition.

Let us, then, start by considering a worry that Clark and Chalmers themselves raise against the Otto case (1998): Despite the functional equilibrium between Otto's notebook and Inga's biological memory in driving their behavior, and despite the satisfaction of the 'glue and trust' criteria by both processes of information retrieval, Clark and Chalmers note that an obvious objection would be to say that all Otto actually believes is that the address is in his notebook, and then describe his actions in two steps: In the 1st step Otto initially believes that the address is in the notebook, *which* leads him to the 2nd step of consulting his notebook and, eventually, to form a new belief about the specific address.

In spite of the initial plausibility of this description, Clark and Chalmers argue (1998) that this is not the right way to go. To see why, they propose, one should treat Inga in the same way: Inga believes that the address is stored in her biological memory, she thinks, consults her memory and finally forms a new belief about the actual address. But, this way of explaining Inga's behavior adds unnecessary complexity, because "Inga relies on no beliefs about her memory as such. She just uses it transparently as it were" (Clark 2010*a*, 46). The same, though, goes for Otto who, having satisfied the 'glue and trust' criteria, is so used to using the notebook that he accesses it automatically. Recall that the satisfaction of these criteria is meant to ensure that the notebook is transparent equipment for Otto just as biological memory is for Inga. In any case, it adds needless and psychologically unreal complexity to introduce additional beliefs about the components of one's cognitive system.

Next, consider the objection from consciousness (Clark & Chalmers 1998, 10). It has been claimed that we should identify the cognitive with the conscious. Consequently, given that it seems implausible to extend consciousness beyond the head, what is external to the organism turns out not to be conscious and, therefore, not cognitive either. To this, however, Clark and Chalmers reply that not every cognitive process is a conscious one. On the contrary, it is a commonplace that many cognitive processes such as memory retrieval, linguistic processes and skill acquisition, to name a few, fall well out the borders of consciousness. So, alone by the speculation that consciousness may be necessarily internal does not follow that what is cognitive must be internal too.

A similar worry raised by Keith Butler (1998) concerns cognitive control. It is beyond doubt, the objection goes, that the locus of computational and cognitive control rests inside the agent's head. Therefore, if cognitive control is an essential feature of the mind, then contributing external processes can be safely excluded from it. Clark, however, responds that this is a false worry and the best way to see why is to readdress the problem to the inner realm; should we not count as part of one's mind any neural subsystems that are not the final arbiters of action and choice? "Suppose only my frontal lobes have the final say—does that shrink the real mind to just the frontal lobes!? What if, as Dennett suspects, no neural subsystem has always and everywhere the 'final say? *Has the mind and self simply disappeared*?"(Clark 2006, 111) But, even if there is some particular inner locus of ultimate choice, we should not identify that with the cognitive agent's mind. My long-term memory is no more an ultimate arbiter of choice than is Otto's notebook. Should I deny my long-term memory as being partially constitutive of my mind? The answer seems to be clearly negative as it would "divorce my identity as an agent from the whole body of memories and dispositional beliefs that guide and shape my behaviors", shrinking the mind and self beyond recognition (Clark 2010*a*, 56). It seems that cognitive control cannot be the mark of the mind.

Another common objection concerns Otto's perceptive rather than introspective route for recalling information. Whereas a normal subject would introspect in order to remember some information, Otto uses vision in order to read the address written in his notebook. Given the fact that the results are obtained in so fundamentally different ways, the explanation of the two cases should not be the same. Clark (2010*a*, 56), however, answers back that whether the recall of information should count as perceptual or introspective depends on how one treats the overall case. From the point of view of the extended cognition, Otto and his notebook count *as a single unified system* and the flow of information is internal to that system. Notice, then, that if introspection is used to describe the internal flow of information within a cognitive system, this seems to be clearly satisfied by the coupled system of Otto and his notebook, no matter that vision, instead of one's inner eye, is involved. If, on the contrary, one argued that the meaning of introspection is to look into one's own head one would, again, beg the question.

So finally, let us consider a kind of worry pertaining to the dissimilarity between the inner and the outer contributions to the cognitive system. Traditional-minded cognitive theorists point to the fact that inner cognitive processes differ so much from the external ones proposed by the HEC theorist that we should not treat them in the same way. Terry Dartnall, for instance, has raised the objection[34] that Otto's notebook cannot count as a form of long-term external memory as it significantly differs from the active nature of biological memory; memory is not a kind of static store of information awaiting retrieval and use. Imagine, Dartnall argues, that you had a chip in your head that gave you access to a treatise in nuclear physics. Could you say that you know about nuclear physics? The answer seems to be negative, because sterile text cannot support cognition; text based storage is so unlike biological memory.

And in a similar vein, Robert Rupert (2004) argues that Otto's way of recalling information is essentially different to biological memory to such an

[34] In personal communication with Clark.

extent that the two information-retrieving mechanisms cannot be both treated as mental processes. Indicatively, Rupert notes, retrieving information from the notebook does not seem likely to exhibit the 'negative transfer' and/or the 'generation' effects which are typically manifested in the process of recalling information from biological memory.[35]

Now, with respect to Dartnall's worry, Clark happily accepts the active nature of biological memory, but his claim is not that the notebook on its own would constitute any kind of cognitive system. "It would not but in this respect it is no worse off than a single neuron or neural population" (2010*a*, 53). Based on the parity principle, the claim was that so long as Otto's employment of the notebook functions in a similar way in driving his behavior as a normal subject's employment of his/her biological memory does, the two processes of information storage and retrieval should be equally treated.[36] Granted, the information stored in the notebook may not affect the rest of Otto's ongoing subconscious parallel cognitive processes in the same way that biological memory does. Think, for example, about "the ongoing underground reorganizations, interpolations, and creative mergers that characterize much of biological memory (Clark 2010*a*, 54). (Still, however, given that we should now consider as the unit of analysis the coupled system of Otto and his notebook, then practical aspects of the two units' interaction may give rise to similar—though of course not identical—effects. And, *mutatis mutandis*, the same could be claimed about the 'negative transfer' and 'generation' effects). Even if, however, the extended system does not have the same (or similar) effects as those its biological counterpart would have to the rest of Otto's mind, "when called upon, its immediate contributions to Otto's behavior still fit the profile of a stored belief" (Clark 2010*a*, 54).

In other words, the parity principle is not an irrational demand for identity between essentially different processes. "Parity is not about the outer performing just like the (human-specific) inner" (Clark 2008, 114); it is a way to intuitively judge what may belong to the process of cognition rather than, say, digestion by not being distracted by the boundaries of skin and skull. In

[35] 'Negative transfer' is a particular form of interference effect, which appears when past learning detrimentally effects the subjects' capacity to learn and remember new associations. The 'generation' effect on the other hand is a mnemonic advantage of subjects who generate their own meaningful connections between pieces of materials learned. For more details see (Rupert 2004).

[36] See fn. 27 again.

this way, we also safeguard ourselves from possibly denying cognition to intelligent aliens whose brains may not be exactly the same as ours. It is a way to avoid the highly chauvinistic thought that what should count as cognitive should bear a fine-grained sameness of processing and storage with the human brain. Similarly then, with respect to Rupert's version of the dissimilarity objection, Clark writes: "just because some alien neural system failed to match our own in various ways (perhaps they fail to exhibit the "generation effect" during recall […]) we should not *thereby* be forced to count the action of such systems as noncognitive" (Clark 2008, 114-5).

To bring this brief overview to an end, then, the above discussion suggests there is no easy or straightforward way to refute HEC. Instead, opponents of HEC must submerge below the surface of ordinary intuitions regarding the domain of the mental—as this is exactly what is called into question—into questions which are closely related to the nature of the mind. The exposition of these questions is the aim of the following section where I discuss a series of interrelated objections that reach—and thereby allow us to clearly delineate—the hardcore of HEC.[37]

2.3) 'Cognitive Bloat', the 'Coupling-Constitution' Fallacy, HEMC, and Dynamical Systems Theory

2.3.1) 'Cognitive Bloat', the 'Coupling-Constitution' Fallacy, and HEMC

Let us now focus on a sequence of interrelated objections that appears to be a

[37] I nowhere explicitly address questions inspired from the field of philosophy of science. Specifically, I do not discuss whether there could be a science of the extended mind in general, or scientific kinds of extended (*aka* hybrid) mental states in particular (see (Clark 1998; 2007; 2008, 2010*a*; 2010*b*), (Rupert 2004; 2009), (Adams and Aizawa 2001; 2008; 2010), (Menary 2006; 2007)). The reason is twofold. First, it seems that said objections are neither relevant to the broader epistemological project I want to argue about, nor will they promote a better understanding of HEC. Second, these objections seem to rest on significantly outdated methodological grounds such as conservatism (see Rupert 2009). Cognitive science, far from having produced a mature paradigm, cannot really dictate the direction of future research. But even if there was a mature research programme available, conservatism is not a well-advised attitude since, as prominent philosophers of science—such as Lakatos (1970)—have observed, when rival alternative research programmes make their appearance, it is necessary to allow sufficient time in order for auxiliary hypotheses and accompanying research and experimental techniques to develop, before start judging whether they are progressive or degenerating. And as it should be clear, HEC is a nascent research programme whose advantages, if any, are yet to be disclosed.

particularly worrying one since it starts by challenging the necessary conditions for HEC and ends up by questioning the very metaphysical core of the view.

Recall the three 'glue and trust' criteria that an external part of the world must satisfy so as to count as a genuine part of one's cognitive economy: (i) its employment must be typical and reliably available, (ii) any information retrieved on its basis must be more-or-less automatically endorsed, and (iii) the information in the resource should be easily accessible as and when required. The problems for the HEC theorist begin with the observation that these criteria may be far too easily satisfied and, therefore, they are not enough to ensure that some external element can count as part of one's cognitive system.[38]

For example, Rupert (2004, 401-5) argues that when a person has access to a phonebook, or a directory service through the use of her cellular phone she can be said to satisfy the criteria that Clark has set forth. Clearly, however, it would be highly counterintuitive to conclude that the phonebook, or the directory service is part of that person's overall cognitive system allowing her to have non-occurent true beliefs about the phone numbers of everyone whose number is listed.

In other words, if any external element that both satisfies the 'glue and trust' criteria and causally affects one's cognitive processes is to count as part of one's cognitive system, we are going to be led to a 'cognitive bloat' (Clark 2001, Rowlands 2009) whereby cognition will seem like leaking all the way out in implausibly many directions. We will be led to an "unacceptable proliferation of systems (many of them extremely short lived)" (Rupert 2004, 396), such that we will loose our grip on the persisting and distinct cognitive agents that are meant to be our objects of study.

Notice, moreover, that this would be the outcome of committing the 'coupling-constitution' fallacy that Adams and Aizawa have pointed out. The objection is that proponents of HEC often put forward their view by arguing

[38] In (Clark & Chalmers 1998, 17) the authors consider a further criterion: "Fourth, the information in the notebook has been consciously endorsed at some point in the past, and indeed is there as a consequence of this endorsement". As the authors further note, however, "the status of the fourth feature as a criterion for belief is arguable (perhaps one can acquire beliefs through subliminal perception, or through memory tampering?)", so they subsequently drop the said criterion.

that since an external process (e.g., the directory service) causally affects a cognitive process (e.g., the search for a phone number), then the external process is a genuine part of the overall cognitive process. But this, Adams and Aizawa note, is fallacious: "it simply does not follow from the fact that process X is in some way causally connected to a cognitive process that X is thereby part of that cognitive process" (2008, 91).[39]

Accordingly, instead of arguing for the constitutive contribution of the external artifacts to one's overall cognitive economy, one should simply endorse the much less provocative idea that cognition is many times *dependent* on external elements. Consequently, one should better opt for a less radical position which has come to be known as the Hypothesis of Embedded Cognition (HEMC) (Rupert 2004, 393).

> HEMC: Cognitive processes depend very heavily, in hitherto unexpected ways, on organismically external props and devices and on the structure of the external environment in which cognition takes place.

This hypothesis is close to HEC as it acknowledges the dependence of cognition on its environment. It is, however, a much more conservative view because it denies that environmental aspects are proper parts of cognition; external factors may only serve as tools and props to cognition, which is restricted within the organismic brain or, at most, the organismic body as a whole. Thus, according to HEMC, cognition is organism-bound, potentially aided by environmental factors, but not extended to them. In other words, HEMC denies that cognitive mechanisms are external, but it also denies that a mechanistic explanation of how psychological processes work should be a purely internal story. "An advocate of HEMC may claim that cognitive mechanisms are internal, but that the mechanistic explanation of how they work is a complex story involving both internal activity and environmental resources" (Sprevak 2010, 356).

More precisely, it has been argued that since both accounts are

[39] Call this the simple version of the fallacy. Adams and Aizawa have also identified a second version of it, which unfolds in two steps: "The first is to move from the observation of some sort of causal connection to the claim that the brain, body and relevant parts of the world form a cognitive system. The second step is a tacit shift from the hypothesis that something constitutes a system to the hypothesis that is an instance of extended cognition" (Adams & Aizawa 2008, 92). This, however, is again fallacious: "It simply does not follow from the fact that one has identified an *X* system in terms of a causal process of type *X* that that process pervades every component of the system" (*ibid.*, 125).

concerned with the way agents interact with their environments, both views will produce the same causal explanations with respect to cases where agents employ artifacts. But, if this is true, then HEMC allegedly wins the day on the basis of conservatism and simplicity: "If two theories embrace structurally equivalent explanations (with or without the same labels), but one of those theories simply tacks on commitment to an additional kind of enitity [e.g., extended cognitive processes, or extended cognitive systems], of no causal significance, then the relative simplicity comparison is straightforward (Rupert 2009, 18).[40]

Should we then abandon HEC on the face of MEMC, or are there reasons for not giving up so easily? Could there be a principled way to individuate systems in general, and cognitive systems in particular, such that we can avoid the 'coupling-constitution' fallacy and the related 'cognitive bloat' worry? Moreover, is it true that HEC and HEMC can provide the same mechanistic explanations and that the only difference between the two is that the former unnecessarily postulates the existence of extended cognitive systems? To answer, we must first focus on dynamical systems theory.

Before moving on, however, let me explain why dynamical systems theory might be relevant here. First, it is the best available tool for modeling and understanding systems that continuously and mutually interact with their environment—which is exactly what extended cognitive systems are supposed to be. Second, within cognitive science, there has lately been an increasing tendency to conceptualize cognitive systems, in general, as dynamical systems that are environmentally (and socially) embedded, and, thereby, best explained using the tools of dynamical systems theory. Take, for example, work in evolutionary robotics (Harvey, Husbands, and Cliff 1994; Husbands, Harvey, and Cliff, 1995; Harvey et al., 1997; Di Paolo 2003; Wheeler 2005), or the work of Lee (2006) who has provided a generic dynamical model that can explain equally well a diverse range of skillful activities such as wing retraction by diving gannets, landing pigeons and humming birds, humans hitting balls, somersaulting, long jumping, putting in golf, and steering. And apart from the modeling of such skillful activities, to which one may nevertheless deny the status of being really cognitive,

[40] Mark Sprevak (2010), however, argues that such a straightforward comparison is, actually, not possible.

dynamical systems theory has also being invoked in order to do significant work on the understanding of unquestionably cognitive phenomena such as language use, decision-making, and social coordination (see, for example, Schmidt, Carello, and Turvey 1990; Busemeyer and Townsend 1995; Roe, Busemeyer, and Townsend 2001; Busemeyer, Townsend, and Stout 2002; Port 2003; Richardson, Marsh, and Schmidt 2005; Oullier et al. 2005; van Orden, Holden, and Turvey 2005; Dale and Spivey 2006; Spivey and Dale 2006; Marsh et al. 2006; Turvey and Moreno 2006; Richardson, Dale, and Kirkham 2007; Richardson et al. 2007; Stephen et al. 2007; McKinstry, Dale and Spivey 2008). And finally, we need not suppose that dynamical systems theory is or will only be relevant to the modeling of cognitive tasks that involve the interaction of the agent with his environment. As recent work by Bressler and Kelso (2001), Thompson and Varela (2001), Varela et al. (2001), Bressler (2002), and Kelso and Engstrøm (2006) demonstrates, dynamical system models work both in brain-only explanations and in brain–body–environment ones. So, having offered this brief note on the scope of dynamical systems theory within cognitive science and how it might be related to our present considerations, let us turn to the theory itself.

2.3.2) Dynamical Systems Theory

One of the primary activities of several scientific disciplines, such as physics, chemistry, biology and the social sciences as well, is the study of systems. Systems are sets of interdependent elements, objects, entities, or items standing in interrelations, on the basis of specific processes they take part in and give rise to, thus forming a unified whole. Of course, an element, object, entity, or item can be part of several systems at the same time, depending on the kind of processes it engages in. Thus, whether some object counts as a component of a system always depends on the phenomenon under study and, more in particular, on the processes that are thought to give rise to the relevant phenomenon.

Similarly, van Gelder writes that "a (concrete) state-dependent system is a set of features, or aspects of the world which change over time interdependently, that is, in such a way that the nature of the change in any member of the system at any given time depends on the state of the members of the system at that time" (1995, 363). Van Gelder further explains that for any two (or more) identical in all relevant physical detail concrete systems (take for example two grand-father clocks) there will be a common abstract structure in their behavior, whose general properties can be studied independently of any particular mechanical device. This is a mathematical structure and is called an abstract state-dependent system. Whereas concrete systems belong to the real world, abstract systems exist only in the timeless and changeless realm of pure mathematical form. Abstract state-dependent systems, however, can be *realized* by particular parts of the real, physical world. "An abstract system is realized by some part of the world when we can systematically classify its states (for example, by measurement) such that the sequences of states the concrete system undergoes is found to replicate the sequences specified by the abstract model" (van Gelder 1995, 364).

Now, when scientists study systems, they often provide models. In general, a model is just another system whose behavior is already better understood—or for some (usually practical) reason, it is more open to exploration—and which is supposed to be similar in relevant respects to the target system. Moreover,

> if a model is sufficiently good, then we suppose that it somehow captures the nature of the explanatory target. What does this mean? Well, if the model is an abstract state-dependent system, then we suppose that the target system realizes the abstract system, or one relevantly like it. If the model is a concrete system, then we suppose that the model and the target system are systems of the same kind, in the sense that they both realize the same abstract system (or relevantly similar systems) (Van Gelder 1995, 364-365).

Now, as far as the scientific study of systems is concerned, on one hand, we have *dynamical modeling,* which is the part of applied mathematics that is concerned with understanding natural phenomena by providing abstract dynamical models for them. *Dynamical Systems Theory* (DST), on the other hand, is a branch of theoretical mathematics, which is concerned with the properties of abstract dynamical models. The general strategy of DST is to conceptualize systems geometrically, in terms of positions, distances, regions and trajectories within the space of a system's possible states. DST is thus primarily concerned with the geometrical properties of the *flow* of the system, which is the entire range of the possible *trajectories* of an abstract dynamical system. Let us engage with some of the technical terms involved in DST in order to get a bit clearer.

In general, every dynamical system is characterized by a set of *state variables x* and a *dynamical law L*, which regulates the change of those *state variables* across time. The set of all possible values of *state variables* is called the system's *state space.* And, if the *dynamical law, L,* depends only on the values of the *state variables* and the values of some set of *fixed parameters u*, then the system is called *autonomous*. This will be the first of three kinds of systems that we will be here concerned with.

Moreover, the *dynamical law, L,* of a continuous-time dynamical system is a set of differential equations x'=*L*(x, u) and defines a *vector field* on the *state space* (here we can imagine the *state space* as a 2-dimensional plane, which depicts in dots all the possible states the system can be in, and the *vector field* as arrows on this plane, indicating the next state the system will be in, given any one of its previous states).[41] Further, starting from some initial state x_0 the law *L* generates a sequence of states, which is called the *trajectory* of the system. The *trajectory* is related to the *vector field* in that its tangent at each point is equal to the value of the vector field at that point (that is, the arrows that represent the *vector field* on the state space, determine the direction of the *trajectory* of the system). The set of all *trajectories* through every point in the *state space* is called the *flow* and, as previously mentioned, DST is primarily interested in the geometrical structure of the entire *flow* of the system; i.e., the geometrical, or topological properties of all the possible behaviors the system might exhibit across time.

Here are some important behaviors that a system may exhibit. Sometimes, a system can converge to *limit sets,* which are the sets of points that are unaffected by the *dynamical law, L,* in that if the state of a dynamical system enters in a limit set, the *dynamical law* will keep it there indefinitely. Of particular interest are the stable *limit sets*, also called *attractors. Attractors* have the interesting property that they gravitate *trajectories* passing through all nearby *states*; if a system's *state* is disturbed away from the *attractor* to a sufficiently small extent, then the *dynamical law* will bring the state back to the *attractor*. Accordingly, the set of initial *states* that converge to a given *attractor*

[41] There are also discrete time systems whose evolution, unlike continuous-time systems, does not unfold continuously across time, but, instead in discrete steps. The *dynamical law* of such systems is simply a map from current state to next state. I will here focus only on continuous-time dynamical systems.

is called its *basin of attraction.* The portions of the *trajectories* that are found within a *basin of attraction*, but which do not lie in the *attractor* itself are termed *transients*. The reason why *attractors* are important is because they govern the long-term behavior of a physical system. Regardless its initial *state,* a physical dynamical system will always settle near an *attractor* after *transients* have passed. In contrast, *repellors* are *limit sets,* which are unstable in that some nearby *trajectories* diverge from them; if the *state* of the system is moved even slightly away from the *repellors,* then the *dynamical law* will carry it away.

In general, the *state space* of dynamical systems contains multiple *repellors* and *attractors* (each surrounded by its own *basin of attraction*), which determine the trajectories the system may take. In other words, those *basins of attraction* constitute the different *phases* the system can enter into, and their graphical representation is called the *phase portrait* of the system.

Now, so far, we have been keeping the *parameters u* of the *dynamical law, L,* constant and have been concerned with the general features of the resulting *flow* (i.e., the set of all possible sequences of *states* (i.e., *trajectories*) the system may exhibit) only as state variables change. Since, however, the *dynamical law* (x′=*L*(x, u)) is also a function of *u*, changes in *u* will definitely bring changes to the resulting *flow*. Even though most dynamical systems are *structurally stable* in that most *parameter* settings will produce small changes in the *flow*,[42] dynamical systems can become *structurally unstable,* such that very small changes in *parameter* values can produce substantial changes in the *flow*, bringing about *phase portraits* which are qualitatively different from the initial one.[43] These qualitative changes in the system's *flow* are called *bifurcations.*

Changing, therefore, the parameters of a system can bring about both quantitative and qualitative changes. What has been said so far, however, concerns *autonomous* dynamical systems, i.e., systems whose *parameters* are held constant for the duration of any particular *trajectory.* Allowing, however, *parameters* to change across time as the *trajectory* unfolds, gives rise to a second type of systems, viz., *nonautonomous systems*. *Nonautonomous* dynamical systems are systems in which one or more *parameters* are allowed to vary in time: x′=*L*(x(t), u(t)). And since "the flow is a function of the

[42] "Limit sets and basins of attraction may deform and move around a bit, but the new flow will be qualitatively similar (i.e., topologically equivalent, or *homeomorphic*) to the old one" (Beer 1995, 180).

[43] For example, new attractors may appear and old ones may disappear.

parameters, in a nonautonomous dynamical system, the system state is governed by a flow which is changing in time (perhaps drastically if the parameter values cross bifurcation points in parameter space) (Beer 1995, 180).[44]

Now, in general, when the *parameters u* remain constant they either refer to one of the internal features of the system that may be manipulated (but which remains fixed during the system's operation), or to the background conditions the system operates in. When the *parameters* change over time, however, we can think of them as *inputs* to the system.

Of particular interest is to see what happens when these changing inputs do not just originate from the system's dynamical environment, but from another system with which the system under study mutually interacts. This mutual interaction gives rise to the third and last type of system we will be here concerned with, namely *coupled* systems. Typical examples of such systems include two mutually interconnected pendulums, the watt governor and a rotation engine, and, possibly, cognitive agents employing epistemic artifacts.

Following Beer (1995), I will here present the case by focusing on an agent and some specific aspect of its environment; i.e., the two nonautonomous dynamical systems whose *dynamical law* is *A* and *E*, respectively. We will also assume that *A* and *E* are continuous-time dynamical systems: $x_A' = A(x_A, u_A)$ and $x_E' = E(x_E, u_E)$. Now, to say that *A* and *E* engage into a constant mutual interaction is to say that the two systems are *coupled nonautonomous* dynamical systems. More specifically, two (*nonautonomous*) systems are *coupled* when the parameters of each system function as some of the *state variables* of the other, and vice versa.[45]

[44] A parameter space is the set of all the values of the parameters encountered in a particular mathematical model.

[45] Important note: In the example Beer offers in his paper, the environmental aspect *E* actually refers to the agent's body, which can be seen as a non-autonomous system that can be coupled to another non-autonomous system (in Beer's example, the agent's neural network). Several authors who have been inspired by Beer's understanding of autonomous agents, however, take *E* to either refer to some ambient feature of the environment (e.g., light), or to some particular object of perception like a tree. The problem, however, is that ambient features of the environment cannot be treated as systems in their own right, and even though objects of perception like trees can be treated as systems in their own right, as far as the agent is concerned, they are *autonomous* systems. Therefore, neither ambient features of the environment nor objects of perception can be seen as *non-autonomous* systems that can be coupled to some *non-autonomous* agent. This is an important note, because by not paying attention to the fact that *coupled systems can only consist of mutually interacting non-autonomous*

Furthermore, we can represent this mutual interaction as a function *S* from the environmental aspect's *state variables* to the agent's *parameters*—such that it captures all the ways in which the environmental aspect under consideration can affect the agent—and a function *M* from the agent's *state variables* to the environmental aspect's *parameters*—such that it includes all the possible ways in which the agent may have an effect on the particular aspect of the environment. Thus $S(x_E)$ represents the effects of the environmental aspect on the agent and $M(x_A)$ represents the effects of the agent on the environmental aspect. We thus have the following *dynamical laws*:

$$X_A' = A\ (x_A, S(x_E), u_A^*)$$

$$X_E' = E\ (x_E, M(x_A), u_E^*)$$

Where u_A^* and u_E^* simply represent any parameters of *A* and *E,* respectively, which are not affected by the coupling.

Now, echoing Beer (182):

> I cannot overemphasize the fundamental role that feedback plays in this relationship. Any action that an agent takes affects its environment in some way through M, which in turn affects the agent itself through the feedback it receives from its environment via S. Likewise, the environment's effects on an agent through S are fed back through M to in turn affect the environment itself. Thus, each of the two dynamical systems is continuously deforming the flow of the other (perhaps drastically if any coupling parameters cross bifurcation points in the receiving system's parameter space) and therefore influencing its subsequent trajectory. […] It is therefore perhaps most accurate to view an agent and its environment as mutual sources of perturbation, with each system continuously influencing the other's potential for subsequent interaction.

Accordingly, given 1) this kind of reciprocal direct dependence between *A* and *E*, and 2) the definition of systems as sets of interdependent elements standing in interrelations on the basis of specific processes they participate in and give rise to, we can view the two *coupled nonautonomous* systems *A* and *E* as a unified *autonomous* dynamical system *U…*

> whose state variables are the union of the state variables of *A* and *E* and whose dynamical laws are given by all the interrelations (including *S* and *M*)

systems, we will first have a wrong understanding of what a coupled system is supposed to be, which will, in turn, lead to wrong and unsurprisingly counterintuitive conclusions.

> among this larger set of state variables and their derivatives. […] Any trajectories observed in the interaction between the nonautonomous dynamical systems *A* and *E* must be trajectories of the larger autonomous dynamical system *U*. Furthermore, after transients have died out, the observed patterns of interactions between *A* and *E* must represent an attractor of *U* (Beer 1995, 183).

We thus see that the *coupling* of *nonautonomous* dynamical systems into one *autonomous* unified system can give rise to a richer behavior than the behavior any individual subsystem could produce on its own, and some of the geometrical properties of the *flow* of the larger system will not be attributable to either subsystem alone. "Therefore, an agent's behavior properly resides only in the dynamics of the coupled system *U* and not in the individual dynamics of either *E* or *A* alone" (*ibid.*, 183).

What this means is that this mutual interaction gives rise to new systemic properties that do not belong to any of the subsystems, but to the ongoing process of interaction which is internal only to the overall coupled system. We also noted that system individuation does not depend on any physical boundaries, but, instead, on the processes one is interested in, and which emerge out of component interactions. So, taking these two last points together provides us with a first reason to think that the (ontological) postulation of *coupled* systems is far from redundant.[46] We can call this the 'systemic properties' argument for the existence of *coupled* systems.

There is, however, yet another reason for which the postulation of *coupled* systems appears to be necessary, and which concerns the nature of the components' interaction involved in such systems. Let me try to spell this out. As noted before, the *law* of *nonautonomous* systems is a function both of the systems' *state variables x*, and its *parameters u*: $x' = L\ (x(t), u(t))$. That is, the overall behavior of the system (i.e., the *flow* of the system) depends on the changing aspects of the system (i.e., the system's *state variables*) and its *parameters*. As far as the parameters are concerned, we noted that when they remain constant for the duration of the system's operation they refer either to

[46] Another way to put the same point could be to claim that postulating a single coupled system in such cases brings explanatory value. That is, the postulation of coupled systems is necessary with respect to the explanation of certain systemic properties, which, otherwise, we would be at a loss how to account for. Coupled systems, therefore, are not open to the common eliminativist line that *X*s do not exist because our best explanations are not committed to the existence of *X*s, i.e., that positing *X*s does no explanatory work. Thanks to Mark Sprevak for pointing out to me this alternative phrasing.

some of the system's internal features (e.g., the material it is made of), or to the stable background conditions the system operates in. When parameters change over time, however, we can think of them as *inputs* to the system. Accordingly, when we have two causally (but not mutually) dependent systems the *input* refers to the effects of the affecting system[47] on the affected system. The *output*, on the other hand, refers to the affected system's reaction (i.e., the system's behavior) to its *input*, but which, by hypothesis (remember there is only one-way dependence), has no substantial direct effects on the affecting system's dynamics. Thus, it will only be represented by quantitative differences in one, or more of the affected system's *state variables.*

Notice, however, that in cases of *nonautonomous coupled* systems, where some of the parameters of each system function as state variables of the other and vice versa, talk of inputs and outputs is not well advised. Indicatively, notice how in the previously quoted passage Beer diligently avoids such talk when he describes a *coupled* system:

> I cannot overemphasize the fundamental role that feedback plays in this relationship. Any action that an agent takes affects its environment in some way through M, which is turn affects the agent itself through the feedback it receives from its environment via S. Likewise, the environment's effects on an agent through S are fed back through M to in turn affect the environment itself.

Apart from Beer's avoidance of such a talk, however, the above quote also suggests why *coupled* systems cannot be described in terms of inputs and outputs. The reason is that the effects of the environment on the agent are *partly defined* by the agent's ongoing activity *at that time*, and *vice versa*. It is, thus, impossible to deconstruct the ongoing causal effects in terms of distinct inputs and outputs from the one system to the other.[48] The feedback loops

[47] The affecting system might be another well-defined system, or the environment in general.

[48] Notice, here, that it is very important to resist the temptation to visualize the case in terms of two components A and E, whereby at time t_0, A is in state x_0, which makes E, at time t_1, enter state x_1, which in turn makes A, at time t_2, enter state x_2, and so on. If this linear story were correct, then we would indeed be able to identify state x_0 as output of A and input to E, and state x_1 as output of E and input to A, and so on. The differential equations, however, describing continuous time dynamical systems refer to infinitesimal differences (hence the name *differential* equations) in time, and so, theoretically, $t_0 = t_1$. Therefore, since x_0 and x_1 are supposed to occur at the same time, there is, again, no way in which they could be conceptualized in terms of inputs and outputs from A to E, and vice versa. More precisely, system A's subsequent state x_2 depends both on system E's state x_1 and its own state x_0 (which affects system E's state x_1) at time $t_0 = t_1$, and so on.

between the two systems give rise to a dense, non-linear interaction, which is in fact *internal* to a larger system comprising of both the agent and the relevant environmental aspect, and which cannot be broken down in terms of distinct inputs and outputs from each subsystem to the other.

Put another way, we might say that in cases where two *nonautonomous* systems mutually interact on the basis of ongoing feedback loops, there is an ongoing *causal amalgam* between the two units that disallows their decomposition into two separate systems on the basis of distinct *inputs* and *outputs.* The reason is that the effects on each component are not exogenous to the component itself and so cannot be properly thought of as *inputs* to the component. Likewise, the way each component affects the other is directly related to the component to be affected and so cannot be properly conceptualized as *output* from the affecting to the affected component. So, again, since we cannot properly disentangle the interactivity of the two components in terms of distinct inputs and outputs from the one to the other, we must accept they constitute an overall system comprising of both of them. We can call this the 'ongoing feedback loops' argument for the (ontological) postulation of *coupled* systems.

So, to close this section, we now have two distinct arguments for the postulation of *coupled* systems. First, the properties of the interaction of *coupled systems* as studied by DST cannot be attributed to any of the contributing systems alone, but must instead be attributed to the overall *coupled* system. Accordingly, we *must* postulate the coupled system. Second, in cases of ongoing feedback loops between *coupled* systems, there is a very dense causal interdependence that disallows us to disentangle them in terms of distinct *inputs* and *outputs* from the one system to the other (the reason being that the effects of each component to the other are not entirely endogenous to the affecting component itself, and vice versa). Accordingly, we *cannot but* postulate the coupled system. Overall, then, we might say that the *constituents* of a system are the interdependent components, which, on the basis of feedback loops, give rise to the processes one is interested in, and which attracted the observer's attention to the relevant components in the first place. Before closing this section, however, let me also note how we can distinguish the overall system from other environmental aspects that may nevertheless affect the system's performance but not constitutively so. For example, sunlight, the ambient temperature, or pressure might affect a system's temporal evolution. However, so long as the system, on its part, has no (measurable) direct effect on those environmental aspects such that they will not, in turn, affect back the system's performance, they can be safely considered as background conditions or inputs with a causal, but not constitutive effect on the system under consideration. Such causal effects will, thus, only be represented in the system's *dynamical law* as its (changing, or constant) *parameters u*.[49]

3.3) Feedback Loops and Cognitive Extension

The discussion of HEC was paused on the interrelated objections of the 'coupling-constitution' fallacy and the 'cognitive bloat' worry. That is, if we take the 'glue and trust' criteria to be both necessary and sufficient for cognitive extension then we are led to 'cognitive bloat', the reason being that the satisfaction of the said criteria cannot ensure that we have not committed the 'coupling-constitution' fallacy. In other words, unless we have a principled way on the basis of which we can decide whether external (to an agent's brain, or organism) components are related to a cognitive process in a merely causal as opposed to constitutive way, 'cognitive bloat' will ensue. Moreover, since 1) no principled account that could draw such a distinction is in sight, and 2) HEMC is a rival hypothesis well suited to provide the same mechanistic explanation as HEC, without making the extra claim that whatever is causally related to a cognitive process is also part of it, thereby avoiding the 'coupling-constitution' fallacy and the concomitant 'cognitive bloat' worry, we should follow HEMC theorists in retaining the term cognitive for the neural, or at most the organismic processes, while fully recognizing the deep ways in which cognition is affected by extraorganismic factors; HEMC is supposed to recognize *all* kinds of cognitive embeddedness, while avoiding the 'cognitive bloat' worry. Cognitive scientists, should, therefore, give up HEC on the face of HEMC.

[49] This apparently generates a problem for the constant parameters that refer to the system's inherent features. Shouldn't the inherent features of a system count as constitutive of the system? Notice, however, that these parameters take their values from the realization basis of the system (i.e., what it is made of). Thus, saying that those parameters of the system are not constitutive of it, really gives rise to the notion of multiple realizability and reveals what are the essential aspects of the system: i.e., the relational properties of the components that arise out of the components' interactions. (This also indicates that parameter spaces can provide information for more specific 'multiple realizability' claims.)

Premise 1) of the previous reasoning, however, now seems questionable. The previous considerations on system individuation provided us with strong support for claiming that when two systems are mutually interdependent on the basis of ongoing feedback loops, then the interactive processes—and their properties—those systems give rise to are internal, and thus belong to, an overall coupled system consisting of both subsystems. In other words, ongoing mutual interdependence on the basis of feedback loops is the criterion by which we can judge whether two seemingly distinct systems *constitute* an overall system, consisting of both of them. Conversely, when no such mutual interaction is in play, but instead a system affects another in a one-way dependence, i.e., the activity of the affected system has no ongoing direct effect on the affecting system—such that no feedback loops are exhibited—then we have a paradigmatic case of a merely causal—as opposed to constitutive—dependence.

So, how is this supposed to help with respect to the 'cognitive bloat' worry? Recall the directory service case again. Even though the agent satisfies the 3 'glue and trust' criteria, no feedback loops between him and the directory service are in play. The person employing those resources does not engage in CRC with them. Those resources are simply causally, as opposed to constitutively, related to the agent in that they deliver her information whose formulation was entirely independent of her. That is, there is no two-way causal interaction between the agent and the epistemic artifact. If an account were to be provided for this case, then it would have to be a linear one, probably in testimonial terms.

Furthermore, using the ongoing mutual interaction on the basis of feedback loops as a criterion of constitution[50] is very much in line with what

[50] Notice that since we have said that systems are individuated on the basis of the processes they give rise to, such that systems may extend in virtue of interactive processes that extend beyond the boundaries of their components, it is not clear which of the two versions of the 'coupling-constitution' fallacy that Adams and Aizawa point to we would commit, in the absence of the 'ongoing feedback loops' criterion of constitution. However, to demonstrate the force of the said criterion, here is how it rules with respect to the two typical examples that Adams and Aizawa use in order to illustrate the two versions of the fallacy.

Simple version: "Consider the expansion of a bimetallic strip in a thermostat. This process is causally linked to the motion of atoms of the air in the room the thermostat is in. The expansion of the strip does not, thereby, become a process that extends into the atoms of the air in the room" (2008, 91). Agreed, but the reason is that the bimetallic strip has no direct effects on the air molecules such that it will in turn affect itself. Of course, Adams and Aizawa could further point out that as time goes by the expansion of the bimetallic strip will turn the heating component off and will thus, in turn, indirectly affect the motion of atoms of the air in the room. Notice, however, that this is not the kind of differential interaction that the has been previously said about HEC. Remember that as we noted earlier it is only when the phenomenon of continuous reciprocal causation (CRC) between the internal and external elements is manifested that we can safely talk about extended cognition.[51] As Clark suggests (2007, sec. 7) "when we confront a recognizably cognitive process, running in some agent, that creates outputs (speech, gesture, expressive movements, written words) that recycled as inputs drive the process along", these inputs should count as constitutive parts of the overall cognitive process.[52] In other words, we should think of extra-organismic parts as constitutive to the overall cognition-generating mechanism when they "emerge as interacting parts of a distributed cognitive engine participating in cognitively potent self-stimulating loops, whose activity is as much an aspect of out thinking as its result" (Clark 2007, sec. 7). And the reason, as it was demonstrated in the previous section, is that in such cases, the outer and the inner contributions come together in a highly

arguments of the previous section from DST are referring to. More precisely, the temperature of the room will only appear as one of the changing parameters, u, in the thermostat's dynamical law, as it is a nonautonomous system.

The Systems version: "The Liquid Freon^{TM} in an older model air conditioning system evaporates in the system's evaporation coil. The evaporator coil, however, is causally linked to such things as a compressor, expansion valve, and air conditioning ductwork. Yet the evaporation does not extend beyond the bounds of the Freon^{TM}. So a process may actively interact with its environment, but this does not mean that it extends to its environment" (*ibid.* 91). Again, agreed; the evaporation process does not extend to the rest of the system. But the whole process of conditioning the air is an overall process whose realization very much depends on the *interaction* between the evaporation coil and the rest of the components of the air conditioning system. Surely, no one would like to identify the air conditioning system with the evaporation coil. To provide a more charitable reading of Adams and Aizawa's objection, what they seem to suggest is that the *essential* process of cooling the air in the whole process of air conditioning does not pervade all aspects of the air conditioning system. In the first place, however, it seems quite suspicious to claim that the evaporation coil is performing the only essential process in air conditioning. Second, in order to draw the parallel between air conditioning and cognitive systems, as Adams and Aizawa wish to do, they need to find the analogy of the 'cooling component' in the case of cognition, such that cognition is restricted within the agent's brain (assuming that the relevant component, or processes can only be found, at least for the time being, within organismic brains). In the absence of such a plausible 'mark of the cognitive', however, when Adams and Aizawa run this version of the 'coupling-constitution' fallacy against the HEC theorist, they either beg the question, or the ball is in their side of the court to convincingly argue for such a 'mark of the cognitive'. See also fn. 54.

[51] Notice that as the following quote suggests , Clark treats CRC only as a sufficient condition on cognitive extension. For the reasons, however, that are being explored in this section, I wish to accentuate its importance by treating it as a necessary condition on cognitive extension.

[52] I must note, here, that given the 'ongoing feedback loops' argument for the ontological postulation of coupled systems (section 3.2) this phrase—literally interpreted—is actually incorrect. That is, properly speaking, when CRC is manifested, one of the points is that there are no distinct inputs and outputs between the interacting systems. Clark probably uses this wording only as a metaphor to indicate the feedback loops generated by the CRC phenomenon. I admit, however, that it is a good way to visualize CRC so I, too, will keep using this phrasing to mean the CRC phenomenon, but, when I do, it must only be interpreted loosely.

complex, probably non-linear, two-way dependence. In such cases, DST treats the two components as a *coupled system* in its own right because "the equation describing the evolution of each component contains a term that factors in the other system's current state (technically, the state variables of the first system are also parameters of the second, and vice versa)" (Clark 2008, 24).[53] Using, thus, the 'feedback loops' criterion to draw systemic boundaries, provides—along with the glue and trust criteria—an objective way for locating cognition without being distracted by the arbitrary boundaries of skin and skull.

So, echoing Hurley (2010), whether external elements might be proper parts of one's cognitive system should not be a metaphysical debate on the nature of the mind, but rather a matter of successful scientific explanation. The internalist should not judge from a pre-scientific and metaphysical standpoint, without independent arguments, whether cognition may be constitutively dependent on external elements, or not, because this is exactly what is at issue here. "To avoid thus begging the question, we should not operate with prior assumptions about where to place the causal-constitutive boundary, but wait on the results of explanation" (Hurley 2010, 106).[54]

[53] For more details see (Van Gelder 1995, 355-8, 373).

[54] Adams and Aizawa (2001; 2008, 2010) claim that the mark of the cognitive is the manipulation of representations with underived content, which is plausibly (at least for the time being) not a feature of any external process, and they thus avoid begging the question against externalism when they put forward the 'coupling-constitution' fallacy.

It is not clear, however, what Adams and Aizawa have in mind and how the said criterion is supposed to promote internalism. One complication is that there is no commonly accepted and non-problematic theory of underived/intrinsic/original content. Second, as Menary (2006) points out, the way Adams and Aizawa make use of the idea that the mark of the cognitive is the manipulation of representations with underived content is either too strict such that it rules out many of the paradigmatic cases of cognition, or too permissive such that it does not disallow cognitive extension.

Finally, after a lot of argumentation, Adams and Aizawa (2010, 70) write: "Clearly we mean that if you have a process that involves no intrinsic content, then the condition rules that the process is non-cognitive". Focusing, however, on the extended cognitive processes which emerge out of the continuous reciprocal causation between extra-neural elements and the organismic brain would surely guarantee the involvement—courtesy of the latter component—of manipulations of representations with underived content, were such entities to exist.

For more details on the debate on underived content as the mark of the cognitive, see (Clark 2008; 2010*b*), (Menary 2006), (Adams and Aizawa 2001; 2008; 2010), (Ross and Ladyman 2010).

Another significant consideration on this topic is that within the dynamical approach to cognition, whether representations exist at all, or what may count as a representation are open questions whose answers might be conceptually alien to the classical computationalist approach that Adams and Aizawa seem to embrace. Indicatively, van Gelder (1995), Pollack (1990), Spivey (2007) and Petitot (1995) have suggested that attractors in the state space of the cognitive system are plausible candidates for bearing representational significance. It is, therefore, an open question whether attractors that arise from the agent's interaction with some artifact could bear non-derived content, and it is by no means clear why we should pre-theoretically exclude such a possibility.

In the framework of the continuously and reciprocally interacting dynamical systems that motivates HEC, the boundaries are not "exogenous to explanatory aims. In cognitive applications, the state space can extend to include dimensions whose variables are bodily and environmental as well as neural, as brain, body and environment interact in mutually shaping patterns" (Hurley 2010, 130). Thus, only time will show whether such externalist approaches to cognition will keep being explanatorily helpful. In any case, however, "the issues between internalism and externalism should be resolved bottom up by such scientific practice, not by advance metaphysics" (Hurley 2010, 107). HEMC and HEC will have to compete with each other on the scientific field.[55]

Before closing this section, however, let me add to the plausibility of the 'ongoing feedback loops' criterion as a necessary condition on HEC, by drawing attention to two arguments in favor of HEC that have appeared in the literature and which, seem to parallel the two arguments that were offered in favor of the postulation of coupled systems in the previous section.

The first argument has been proposed by Clark who notes that, contrary to HEMC, the properties of the extended system are not identical to the properties of its parts. As Clark (2008, 116) writes:

> A further reason to resist the easy assimilation of HEC into HEMC concerns the nature of the interactions between the internal and the external resources themselves. Such interactions, it is important to notice, may be highly complex, nested, and non-linear. As a result there may, in some cases, be no viable means of understanding the behavior and potential of the extended cognitive ensembles by piecemeal decomposition and additive reassembly. To understand the integrated operation of the extended thinking system created, for example, by combining pen, paper, graphics programs, and a trained mathematical brain, it may be quite insufficient to attempt to understand and then combine (!) the properties of pens, papers, graphics programs, and brains.

[55] For an objection to the claim that scientific practice can resolve this debate see (Sprevak 2010). In so arguing, Sprevak has in mind that both HEC and HEMC will produce the same causal explanations—since they are both interested in the interaction of the agent with the her environment—and so scientific explanation will not be able to produce a clear verdict regarding the success of HEC over HEMC, or the other way around. My reasons for thinking that it can are quite different. Should there be cases where the accomplishment of some cognitive task is the product of the mutual interaction between the agent and some artifact, then given the DST framework and the concept of coupled systems (as well as the absence of a plausible mark of the cognitive) any theorist will have to conclude that there is a coupled cognitive system that operates on the basis of extended cognitive processes.

Therefore, in many cases, the extra neural elements and the organismic brain should not be considered—as HEMC, in effect, only allows—as distinct systems whose coupled performance could be linearly analyzed. Instead, the only possible way to study their properties will be as one unified system.[56] But, now, moving back to the 'systemic properties' argument for the existence of coupled systems, notice how it basically conveys the same idea.

Argument number two comes from Francisco Varela who writes: "If one says there is a machine M in which there is a feedback loop through the environment, so that the effects of its output effects its input [i.e., M reciprocally interacts with some environmental aspect through self-stimulating loops], one is in fact talking about a larger system M′ which includes the environment and the feedback loop in its defining organization" ((Varela 1979, 158); quoted by (Hurley 1998, 104). The reason is that the effects of the environment on M are partly defined by M's ongoing activity *and vice versa.* Notice again how this is the same reasoning as the 'ongoing feedback loops' argument for the postulation of coupled systems.

Moreover, it may be claimed that these two arguments, both of which can be run with respect to abstract systems and concrete cognitive systems, should perhaps not be seen as distinct, for they both point towards the same underlying idea. When a behavior emerges out of the interaction of two components, this behavior can only belong to the overall coupled/extended system—as it cannot be broken down in terms of distinct inputs and outputs between components—and its properties cannot be understood without appealing to the said interaction, which again—quite trivially—depends on

[56] It has been pointed out to me that the HEMC theorist may be happy to accept that an agent might be in a continuous and reciprocal causal relation with some aspect in the environment, but still deny cognitive extension. However, this seems not to be an option for the HEMC theorist because given the conceptual framework of Dynamic Systems Theory, in such cases, HEMC collapses into HEC. In fact, Rupert, in chapter 7 of his book (2009), concedes that Dynamical Systems Theory can provide strong support to the hypothesis of extended cognition in just the way I have been here describing (Rupert 2009, 131-4). It is telling that none of the dynamic models that he considers in favor of HEMC concerns a two-way interaction between the organism and some particular environmental aspect. For more details see (Rupert 2009, 137-149). For more on HEMC's failure of aggregativity see Theiner (*manuscript*). Michael Wheeler is one more theorist who has claimed that CRC between brain, body, and environment is a powerful indication of extended cognition. More specifically he writes (2005, 265): "[W]here the complex channels of continuous reciprocal causation cross back and forth over the physical boundaries of skull and skin, the cognitive scientist, operating from a functional or organizational perspective, may face not (i) a brain-body-environment system in which brain, body, and environment form identifiable and isolable subsystems, each of which contributes in a nontrivial way to adaptive success, but rather (ii) just one big system whose capacity to produce adaptive behavior must be understood in an holistic manner".

both subsystems at the same time.

Consequently, then, apart from the 3 'glue and trust' criteria, in order to account for the constitutive status of external elements within one's overall cognitive mechanism, we also need the phenomenon of continuous reciprocal causation (which arises on the basis of ongoing feedback loops) between the outer and the inner parts to take place. These 3+1 criteria seem to jointly ensure the integration of the external artifacts within one's overall cognitive mechanism, thereby overcoming the 'coupling-constitution' fallacy and the 'cognitive bloat' worry. In such cases, the ongoing interaction deeply affects the 'tendencies', or even causes bifurcations in the agent's and/or the artifact's phase portrait, thereby giving rise to behavior, which is distinctive of the relevant extended process.

Finally, as far as the distinction between HEC and HEMC is concerned, what the coupling arguments from DST demonstrate is that cognition is indeed embedded à la HEMC whenever the environment affects the agent in a one-way dependence. However, if and whenever, one's organismic cognitive apparatus interacts with the environment on the basis of ongoing feedback loops, then the HEMC theorist must recognize that the mind is not only embedded, but also extended à la HEC.

2.4) Gestures: An Example of Extended Cognition

Finally, a useful way to illustrate how extended cognition might be realized through the satisfaction of the above conditions is the interesting example of the active role that bodily gestures play in driving thought and reason.

In an extensive attempt to understand the nature of human gesture, the psychologist Goldin-Meadow wonders whether gesture simply serves the purpose of communicating thoughts, or might it be a part of the actual process of thinking (2003)? A list of indicative facts that could suggest abandoning the first alternative is the following: we gesture when talking on the phone, we do it when talking to ourselves, we do it in the dark, gesturing increases with task difficulty, gesture increases when one has to choose between options, gesturing increases when reasoning about a problem rather than merely describing the problem or a known solution (Clark 2007, sec. 5)

One could, however, try to explain most of the above hints by holding that gesturing without a viewer merely demonstrates our habit to gesture in the rest of the normal communicative cases. But, this line of answer is in trouble explaining the additional fact that speakers blind from birth do gesture, and they even do so when speaking to people they know to be blind too.

In order to investigate the impact of restricting gesture from the mix of available resources on thought, Goldin-Meadow asked two groups of children to memorize a list and then solve a mathematical problem before trying to recall the list. One group was allowed to gesture during the problem-solving task, whereas the other group was asked not to. The outcome was that removing gesture during problem-solving had a detrimental effect on the subsequent task of recall. The best explanation, according to Goldin-Meadow, is that gesturing shifts or lightens aspects of the neural cognitive resources, thus freeing up resources for the memory task.

A possible objection to this conclusion, however, is that gesturing does not lighten the neural burden carried out by the free-to-gesture group. Rather, asking the other group not to gesture could have added load to their neural resources, because they had to remember not to gesture. This could equally explain why the free-to-gesture group performed better in the recall task. As luck would have it, however, some subjects from the free-to-gesture group willingly chose not to gesture (thus not having to remember to hold back their natural inclination to gesture), which, nevertheless, turned out to have the same detrimental effect regarding the recall task as with the non-gesture group. This lucky event safeguards the initial explanation from the alternative one; gestures themselves seem to play some active cognitive role.

As Goldin-Meadow writes, "gesture…expands the set of representational tools available to speakers and listeners. It can redundantly reflect information represented through verbal formats or it can augment that information, adding nuances possible only through visual or motor formats" ((Goldin-Meadow 2003, 186); quoted by (Clark 2007, sec. 5)).

In other words, verbal thinking continuously informs and alters gesture, which continuously informs and alters verbal thinking, i.e., the two form a genuinely coupled system. In the case of gesture, therefore, we explicitly observe a cognitive process whose realization outstrips the boundaries of the neural realm, extending to the biological organism.

Crucially, however, we need not suppose that this kind of extension is limited to the purely biological organism for, as Clark notes, the case of gesture bears a lot of similarities and is closely akin with environmentally extended cases such as when one is busy writing and thinking at the same time; "it is not always that fully formed thoughts get committed to paper. Rather, the paper provides a medium in which, this time via some kind of coupled neural-scribbling-reading unfolding, we are enabled to explore ways of thinking that might otherwise be unavailable to us" (Clark 2007, sec. 5).

2.5) Conclusion

To sum up, the hypothesis of extended cognition is primarily motivated by the observation that, in many cases, environmental factors that are external to our biological organism and particularly to our neural system have a large impact on cognition. The most promising rival theory, namely the hypothesis of embedded cognition, seems too weak to explain all the ways in which environmental factors may contribute to the cognitive system. The reason is that HEMC treats the relation between the outer and the inner contributions merely as causal; the intelligent organismic agent consciously chooses to offload bits of his ongoing work to the environment and then reload it, benefiting from this process that can be linearly analyzed. This description, however, does not seem to fit entirely the picture, because the cases which HEC is interested in are cases where the agent is actively coupled to some extra-neural element. This means that the relation between the outer and the inner contributions is interactive in a highly complex non-linear way. And this emerging interaction between the organismic and the extra-neural parts is constitutive of an overall, extended cognitive system and not merely causal, because it not only helps the agent but, literally, drives and constrains her own evolving process of thought and reason.

Moreover, we noted that the satisfaction of the 3+1 criteria (i.e., the three 'glue and trust' criteria along with the phenomenon of continuous reciprocal causation) seems to safeguard the proponent of HEC from the 'coupling-constitution' fallacy and the 'cognitive bloat' worry. We, therefore, do not need to worry that if we embrace HEC we will be led to an "unacceptable proliferation of systems (many of them extremely short lived)" (Rupert 2004, 396), such that we will loose our grip on the persisting and distinct cognitive agents that are meant to be our objects of study. The reason for this, as we noted, is that an external element will not count as properly integrated within one's cognitive system unless it is also reciprocally coupled to the agent's organismic cognitive faculties, and this is only rarely the case. In other words, in order for an artifact to be a constitutive element of one's ongoing cognitive loops, the agent must deliver on its basis outputs, which recycled as inputs will drive his/her cognitive system along. And this does not happen every time the environment has an effect on the agent, but only when the latter actively engages in a continuous interaction with a specific aspect of the former.

Furthermore, this realization also demonstrates that HEC clearly acknowledges the central role of the persisting biological organism in recruiting and maintaining the extended organization in order to eventually accomplish its very own cognizing: "Human cognitive processing (sometimes) extends to the environment surrounding the organism. But, the organism (and within the organism the brain/CNS) remains the core and currently the most active element. Cognition is organism centered even when it is not organism-bound" (Clark 2007, sec. 9).

So, HEC is in no danger of undermining the central role of the persisting biological cognitive agent. Instead, what HEC tries to underlie is that as soon as the extended organization has been formed, then there is no reason to point to the boarders of skin and skull so as to explain the way information flows and is processed, because "once such organization is in place, it is the flow and transformation in an extended, distributed system that provides the actual machinery of ongoing thought and reason" (Clark 2007, sec. 4).

So, having by now seen how cognition may be extended, we should now return to the epistemological discussion and particularly to the intuition that knowledge must be the product of cognitive ability in order to see how these two ideas may be fruitfully combined.

CHAPTER 3

$COGA_{weak}$ and the Hypothesis of Extended Cognition

3.1) Introduction

Having introduced $COGA_{weak}$ and HEC, I would now like to focus on the observation that these two views bear some interesting core similarities, which render them mutually supportive and informative. Moreover, considering the two frameworks side by side will help us better understand the notions of one's 'cognitive character', 'cognitive agency' and 'belief-forming processes' which are central to the formulation of $COGA_{weak}$, but which have so far been rather vaguely discussed.

I shall first outline some interesting parallels that can be drawn between the core concepts and claims of the two views. Once the common points of $COGA_{weak}$ and HEC become apparent, we will have gained an understanding of the epistemological concepts involved in the former, which we shall test against a number of intuitions that spring from several versions of two core thought experiments, namely the *Temp* and the *Alvin* case. Finally, I shall explore a series of extra-organismic knowledge-conducive belief-forming processes, whose employment may extend our cognitive characters beyond our organismic cognitive faculties, including the interesting case of scientific theories. Before moving on, however, here is a prefiguration of some of the main ideas that I will be later expanding upon.

Remember that, according to HEC, "individual cognizing is organism centered even if it is not organism-bound" (Clark 2007, sec. 4). That is, our cognition may be extended beyond our organismic cognitive faculties of the brain/CNS, which, nevertheless, make us the persisting and distinct cognitive agents that we are. In so extending her cognition, the agent incorporates to her cognitive loops extra-neural environmental elements. These elements, though external to her organismic cognitive faculties, form a proper part of the agent's cognitive system as they actively drive her cognition in a highly complex and interactive way. In this way, the agent's cognitive economy can extend beyond her organismic cognitive faculties even though it is centered to them; recall that it is the agent's brain/CNS that is responsible for the recruitment of the extra-neural elements.

Now, returning to COGA$_{\text{weak}}$, it appears that very similar claims can be made on its basis with regards to the cognitive resources of *epistemic agents*. According to COGA$_{\text{weak}}$, an epistemic agent *S* can come to know the truth of some proposition *p* on the basis of some reliable belief-forming process so long as the relevant belief-forming process is appropriately *integrated* within his cognitive character. Interestingly, COGA$_{\text{weak}}$ makes no specifications about whether the process must be wholly realized within the agent's organism or not; in principle, it allows the epistemic agent to extend his cognitive character beyond his cognitive agency by appropriately integrating within the former belief-forming processes that are external to the latter (just as HEC allows a person *S* to extend his cognitive system beyond his organismic cognitive abilities by incorporating to his cognitive loops extra-bodily elements). It thereby seems we have a quite interesting analogue between an agent's potentially extended cognitive system and the epistemological notion of one's cognitive character.

Moreover, think about the notion of one's *cognitive agency* to which, according to COGA$_{\text{weak}}$, the cognitive success must be significantly creditable. The underlying reason for this requirement seems to be that if we want to attribute knowledge to a specific individual *S*, this knowledge should have its origins, at least to a significant degree, in what makes *that* individual *S* a distinct and persisting epistemic agent. But, since according to virtue reliabilism, knowledge must be the product of cognitive ability in a way that respects the *ability intuition* on knowledge, one might conclude that what makes a specific individual *S* a distinct and persisting epistemic agent coincides with what makes *S* a distinct and persisting cognitive agent, i.e., *S*'s cognitive agency.

Note, furthermore, that according to most, if not all, theories of cognition—including HEC—what makes an individual *S* a distinct and persisting *cognitive* agent is her cognitive faculties of the brain/CNS.[57] Consequently, one might conclude, the epistemic agent's cognitive agency is made up of her organismic cognitive faculties of the brain/CNS. Is this

[57] Obviously, I here exclude any dualist alternatives.

conclusion helpful? It appears so, because, in accordance to intuition, it also explains why when the agent comes to know on the basis of a reliable belief-forming process which is for the most part extra-organismic, the agent's cognitive success is still significantly creditable to her cognitive agency: it is her organismic cognitive faculties of the brain/CNS that are responsible for integrating within her cognitive character the partly external belief-forming processes that led her to come to know the truth of the target proposition.

Briefly then, what the above discussion is meant to foreshadow is the fact that two very similar claims can be made on behalf of the two theories under consideration. According to HEC, one's cognitive loops that make up one's cognitive system are centered around one's organismic cognitive faculties of the brain/CNS, but not bound to them as one's cognitive system can extend so as to incorporate extra-neural elements. Similarly, according to COGA$_{weak}$, the epistemic agent's belief-forming processes that make up her cognitive character are centered around her organismic cognitive faculties—that make up her cognitive agency—but are not bound to them as her cognitive character can extend so as to incorporate extra-bodily belief-forming processes.

In other words, straight parallels could be drawn between (i) the agent's cognitive character and the agent's cognitive system, (ii) the partly external belief-forming processes and the extended cognitive loops, and finally (iii) the epistemic agent's cognitive agency and the agent's organismic cognitive faculties, i.e., his brain/CNS, which are responsible for the recruitment of the reliable belief-forming processes. Are these mere analogues or should we think that they are superficially different terms, which, nevertheless, refer to the same notions shared by both COGA$_{weak}$ and HEC?

3.2) COGA$_{weak}$ and the Extended Cognition Hypothesis

In order to answer the above question we should first refresh our memory with respect to the criteria that a belief-forming process must meet so as to deliver knowledge, according to the broader virtue reliabilistic framework.[58]

[58] Notice that since COGA$_{weak}$ is only a necessary condition on knowledge, the following criteria are supposed to be necessary but not sufficient for knowledge. Remember that Pritchard argues that for a full account of knowledge, COGA$_{weak}$ must be supplemented by

In general, the process must be a cognitive ability. In order for this to be so, we noted in section 1.3, first, the process must be a cognitive process. This will ensure that the direction of fit between the belief and the fact will be the correct one. Second, we want the belief-forming process to be reliable, where a reliable process is one that tends to produce true rather than false beliefs. Recall, however, that according to reliabilism, one does not need one's evidence to be necessarily reliable; if forming a belief on a certain kind of evidence constitutes a reliable belief-forming process, it does not matter that one's evidence is only contingently reliable. What this means, according to virtue reliabilism, is that the agent, on his part, does not need to know or have any beliefs about the reliability of his belief-forming process. It is sufficient that he is sensitive to the reliability of his way of forming beliefs simply by it being one of the cognitive processes that constitute his cognitive character, which he employs when he is thinking conscientiously. But as we further noted, in order for a process to be, in principle, part of the agent's cognitive character it must be neither strange nor fleeting. It must not be strange so that he won't reject it when conscientious. And it must be a disposition or a habit of the agent, because it is only dispositions or habits that one can become aware they are unreliable in certain circumstances, and so—without relying on any beliefs about their reliability—be, in principle, able to employ them conscientiously in the rest of the circumstances. And, finally, COGA$_{\text{weak}}$ demands that the process be appropriately integrated within one's cognitive character. This will guarantee both that the relevant belief-forming process is a cognitive process, and that it is indeed part of the agent's cognitive character such that he will, in fact, be conscientious in employing it.[59] So, putting all the above points together: *a belief-forming process counts as a cognitive ability such that it is knowledge-conducive only if it is a reliable belief-forming disposition properly integrated within the agent's cognitive character*. Now, let us recall what the three criteria to be met by non-biological candidates for inclusion into an individual's cognitive system are, as offered by Clark (2010*a*, 46):

1) "That the resource be reliably available and typically invoked". That

the safety condition. Not everyone agrees with this, however, and this is why I write "according to the broader virtue reliabilistic framework".

[59] If the process is not actually integrated within the agent's cognitive character, then we may suppose that the agent won't have become aware when it is inappropriate to employ it, and so he won't be conscientious in employing it, even when he does so in the right circumstances.

is, the agent should habitually and easily invoke the external resource. In other words, its employment must be a *disposition/habit* of the agent's overall cognitive mechanism.

2)"That any information thus retrieved be more-or-less automatically endorsed. It should not usually be subject to critical scrutiny. […] It should be deemed about as trustworthy as something retrieved clearly from biological memory". That is, the information in the resource must be *regarded* as reliable, and not be necessarily reliable, so long as its employment results into an equally trustworthy belief-forming process as the one of forming beliefs on the basis of one's own biological memory.[60]

One might object, however, that being reliable is not the same as being *trustworthy* (i.e., being regarded as reliable).[61] In response, notice first that Clark identifies the notion of trustworthiness of a process with the idea of being "more-or-less automatically endorsed" or, in other words, "not usually being subject to critical scrutiny". What this means is that the target process must have not been (for the most part) problematic in the past. Moreover, the processes under consideration are supposed to be cognitive dispositions/habits of the agent, which he has *repeatedly* employed in the past, and so had they been problematic the agent would have noticed that and responded appropriately. Accordingly, a *trustworthy* belief-forming process, in Clark's account, will be one that tends to produce true rather than false beliefs, which is to say that it will be objectively reliable in the virtue reliabilist's sense. What the agent will deem reliable will be what is objectively reliable, i.e., that which has *not* been (for the most part) problematic in the past.[62]

[60] That is, the process does not need to be, due to underlying logical or quasi-logical relations, 100% reliable. Notice that memory is supposed to be reliable even though one may misremember.

[61] The idea here is that trustworthiness (i.e., being regarded as reliable) might sometimes supervene on values other than objective reliability (i.e., the tendency to lead to success rather than failure). See the next footnote for a description of just such a case.

[62] To elaborate a bit more on this point, it has been objected to me that in the United States, there are many individuals who trust a particular cable TV "news organization", but this news organization often provides misleading information. Viewers of this cable TV "news" channel trust this medium, but it is not reliable in the epistemologist's sense. So, it may now appear that Clark's conditions and the epistemologist's reliability condition come apart. In response, let me draw your attention to two points. First, as noted above, Clark's claim that in order for something to be trustworthy it must not usually be subject to critical scrutiny implies that the agent *would* check the object of his trust on the face of discrepancies. In contrast, the viewers of the American TV news channel appear that they *would not,* and this is a kind of *blind trust* that is quite different from the kind of trust Clark has in mind. Accordingly, the fact that there might be kinds of trust that have nothing to do with objective

Now, notice this negative way of deeming processes reliable with which Clark concurs (i.e., that a trustworthy process is one that is *not usually* subject to critical scrutiny such that it is more-or-less automatically endorsed). What this means is that the agent does not need to have any beliefs about why or whether the belief-forming process is trustworthy; it suffices that it has not repeatedly caught his negative attention in the past. And this seems to be in good agreement with Greco's demand that one's subjective justification need not rely on any beliefs but simply on one's motivation to believe the truth. For example, one may trust one's vision in appropriate circumstances, just because vision has not been notably problematic in the past (in those circumstances). It suffices that one is motivated to believe the truth and will thereby employ the belief-forming process that has not in the past (generally) failed to be conducive towards this end, and, crucially, one will do so without even thinking about it. This methodological point, in turn, appears to be in line with the spirit of the glue and trust criteria which, after all, are meant to ensure the effect of transparent equipment: equipment that we are so fluent and familiar with that we have no beliefs about it in use.

3) "That information contained in the resource should be easily accessible as and when required". That is, the agent must be able to directly and easily access the resource whenever necessary; he must be able to employ it as if it was part of his organismic cognitive mechanism. In other words, the resource must have been *integrated* within the agent's overall cognitive mechanism.

We thus see that in order for non-biological elements to be included into one's cognitive system, their deployment must meet the same criteria as the ones that the belief-forming processes must satisfy so as to count as knowledge-conducive (i.e., to be a cognitive ability). We also need, however,

reliability should not generate any problems for the identification of Clark's 'trustworthiness' condition with the epistemologist's objective reliability criterion.

Second, even if on Clark's account, the American TV "news" channel were to count as part of one's cognitive economy—were it to also satisfy the rest of the criteria for cognitive extension—one could note that this is an issue about which epistemology and philosophy of mind can, and should exchange normative considerations. That is, even if, when philosophizing about what may count as part of an agent's mind, we do not focus on the nature of a good mind, epistemology could point out that a *conscientious* mind should try to believe what is true and thereby employ the resources, which are reliable in the epistemologist's sense. Nevertheless, Clark seems to accommodate this normative dimension of the mind through his "not-usually-subject-to-critical-scrutiny" understanding of trustworthiness, thus having been in line with the virtue reliabilist tradition in epistemology, all along.

to make a further remark concerning the *appropriate* integration of the external elements within an agent's cognitive character.

According to Pritchard, as noted in section 1.4.2, one way to test whether a belief-forming process is properly integrated within one's cognitive character is to consider whether the cognitive success of believing the truth is to a significant degree creditable to the agent's cognitive agency (Pritchard 2010*b*, en. 7). This proposal is based upon the main idea that since COGA$_{weak}$ is a formulation of the ability intuition on knowledge, then so long as one's true belief is significantly creditable to one's cognitive agency, the process by which one came to form one's true belief can be said to have been appropriately integrated within the rest of one's cognitive character, and thereby count as a *bona fide* cognitive ability. Although this might generally be a good way to start judging whether a process can count as a *bona fide* cognitive ability, note that there might be cases where the cognitive success might be significantly creditable to one's cognitive agency even if it is not the product of one's cognitive abilities—we shall come across such a counterexample in the next section where I discuss several thought experiments. It is, therefore, important to find an additional, descriptive criterion to account for what it takes for the belief-forming process to be appropriately integrated within one's cognitive character, such that—and in accordance to COGA$_{weak}$—the cognitive success will be significantly creditable to one's cognitive agency. But where can we find such a descriptive criterion?

As we noted in the previous section, an artifact will not count as properly integrated within one's cognitive system unless it is also reciprocally coupled to the agent's organismic cognitive faculties. That is, in order for an artifact to be a constitutive element of one's ongoing cognitive loops, the agent must deliver on its basis outputs which recycled as inputs will drive his cognitive character along. Therefore, in order for a belief-forming process to be appropriately integrated within one's cognitive character, the phenomenon of continuous reciprocal causation (CRC) must be manifested between the target process and one's organismic cognitive faculties.[63]

The underlying reason for this is that HEC acknowledges the central role of the persisting biological organism in recruiting and maintaining the

[63] Similarly, Greco has noted that "in general, it would seem, cognitive integration is a function of cooperation and interaction, or cooperative interaction, with other aspects of the cognitive system" (2010, 152).

extended organization in order to eventually accomplish its very own cognizing: "Human cognitive processing (sometimes) extends to the environment surrounding the organism. But the organism (and within the organism the brain/CNS) remains the core and currently the most active element. Cognition is organism-centered even when it is not organism-bound" (Clark 2007, sec. 9). Notice, then, that using the phenomenon of CRC to judge whether an artifact has been appropriately integrated within one's cognitive character is in line with Pritchard's suggestion to consider whether the cognitive success will eventually be significantly creditable to one's cognitive agency; it is the agent's organismic cognitive faculties (i.e., the brain/CNS) that are first and foremost responsible for the recruitment of the external elements on whose basis the agent will deliver outputs which recycled as inputs will drive her cognitive character further along, so as to eventually form a true belief with respect to some proposition *p*.

We thus see that the glue and trust criteria as well as the CRC phenomenon (i.e., the 3+1 criteria) are common currency for both $\text{COGA}_{\text{weak}}$ and HEC, thereby rendering both accounts complementary. Reasonably, then, it may also be claimed that there is no principled theoretical bar disallowing extended belief-forming processes to count as knowledge-conducive cognitive abilities. That is, provided the satisfaction of the 3+1 criteria, one's cognitive character may extend beyond the organismic cognitive abilities that make up one's cognitive agency, by incorporating epistemic artifacts. Put differently, given the right conditions, the deployment of epistemic artifacts can count as knowledge-conducive cognitive abilities, allowing us to gain knowledge on their basis in accordance with the ability intuition on knowledge.[64]

3.3) Case studies

So let us now see how the 3+1 criteria rule with respect to several versions of two core thought experiments that Pritchard discusses in (2010*b*), namely the Temp and Alvin case.

[64] Remember, in section 1.4.2 we noted that as virtue reliabilism is usually formulated, it is in trouble explaining how epistemic artifacts can count as proper parts of one's cognitive character. It is only when we embrace Pritchard's virtue reliabilistic $\text{COGA}_{\text{weak}}$ that we can account for such an extended cognitive character.

Temp[65]

> Temp's job is to keep a record of the temperature in the room that he is in. He does this by consulting a thermometer on the wall. As it happens, this way of forming his beliefs about the temperature in the room will always result into a true belief. The reason for this, however, is not because the thermometer is working properly, since in fact it isn't—it is fluctuating randomly within a given range. Crucially, however, there is someone hidden in the room next to the thermostat who, unbeknownst to Temp, makes sure that every time Temp consults the thermometer the temperature in the room is adjusted so that it corresponds to the reading on the thermometer.

Intuitively and according to COGA$_{\text{weak}}$ Temp lacks knowledge of the temperature. As Pritchard explains, even though the way in which Temp forms his true beliefs is reliable, it in no way reflects his cognitive abilities. Therefore, since Temp fails to meet the ability intuition on knowledge, COGA$_{\text{weak}}$ disallows knowledge to Temp. One may further argue that the reason why Temp fails to meet the ability intuition is that the relevant belief-forming process lies outside his head. Consider, however, Alvin.

Alvin[66]

> Alvin has recently developed an unusual brain lesion, a guaranteed side effect of which is that it prompts him to randomly, but reliably, form true beliefs about the product of fairly complicated arithmetical sums.

Alvin's belief-forming process is indeed reliable. However, even though it lies under Alvin's skin, Pritchard argues that Alvin lacks knowledge of the mathematical propositions. The reason, Pritchard claims, is that as in the Temp case, Alvin's cognitive success has nothing to do with Alvin's cognitive abilities but is rather the "fortunate consequence of the otherwise unfortunate fact that he has a brain lesion" (2010*b*, 136). I agree. One, however, may fairly wonder: why think that the way Alvin forms his beliefs is not one of his cognitive abilities? After all, it is part of his brain. In response, we could now claim that Alvin's belief-forming process fails to satisfy the first and the second 'glue and trust' criteria. That is, Alvin's belief-forming process 1) is not

[65] (Pritchard 2009, 48)

[66] Pritchard (2010*b*, 136). The case is adapted from one offered by Plantinga (1993).

one of his dispositions (i.e., one of his habitual cognitive routines) and thereby 2) it cannot have been deemed trustworthy.[67]

Next, Pritchard moves back to the Temp case and asks how would our intuitions change if Temp came to know what is the true source of the reliability of his belief-forming process and that it is reliable. Pritchard (2010*b*, 138) argues that it would make a great difference because in becoming aware of the source of the reliability Temp can now take cognitive responsibility for this cognitive success. Accordingly, his cognitive success is primarily creditable to his cognitive agency and thereby, by Pritchard's suggestion,[68] the relevant belief-forming process (including the broken thermometer, the thermostat, and the hidden helper) has been appropriately integrated within his cognitive character. Temp, therefore, can gain knowledge in this way.

At this point, however, the 3+1 criteria and the results of Pritchard's suggestion to regard a belief-forming process as appropriately integrated within one's cognitive character when the cognitive success is significantly creditable to one's cognitive agency come apart. For in the way Temp forms his beliefs, no continuous reciprocal causation is being observed between Temp and the thermometer on the wall. Instead, the thermometer delivers information to Temp in a one-way causal—as opposed to constitutive—dependence. Notice that I am not saying that Temp cannot gain knowledge when he knows the true source of the reliability of his belief-forming process and that it is reliable. He can. I am rather saying that this is not a case of extended cognition. Instead, a more appropriate epistemic description of this case would be in testimonial terms, whereby Temp has positive reasons to accept and/or no undefeated defeaters to deny the reading of the thermometer.[69]

Next, however, Pritchard (2010*b*, 138) asks the same question about Alvin: what if Alvin comes to know both what the true source of the

[67] Notice that Alvin's lesion is not one of the belief-forming processes he would employ were he motivated to believe the truth. Or, in other words, forming beliefs on the basis of the brain lesion cannot be deemed trustworthy. Alvin cannot automatically endorse the products of this process; it is not the case that, from Alvin's 'point of view', the resource is not *usually* subject to critical scrutiny, simply because it is a recently acquired one.

[68] Recall that, according to Pritchard, when some cognitive success is significantly creditable to one's cognitive agency, then the belief-forming process by which the belief was acquired can be said to have been appropriately integrated within one's cognitive character.

[69] Notice, then, that while it is the case that whenever CRC takes place the cognitive success will be significantly creditable to one's cognitive agency, things do not work the other way around.

reliability of the belief-forming process is and that it is reliable? Pritchard claims that now Alvin can gain knowledge. I agree, because in so being aware of the facts he can now satisfy the first and second 'glue and trust' criteria.[70] Moreover, it would be very implausible to deny CRC in this case. Alvin's lesion is to be found within his brain and so, most likely, the affected area interacts very densely with the other parts of his brain in order to produce the relevant outputs. Becoming, however, aware of the source of the reliability of a belief-forming process is a very strong condition on knowledge and it is exactly what externalist approaches such as virtue reliabilism set out to resist. Accordingly, Pritchard goes on to further explore whether one's belief-forming processes can be integrated within one's cognitive character in weaker ways.

According to Pritchard (2010*b*, 146), one factor that seems to play a central role regarding our intuitions on the integration of a belief-forming process is whether it has always been present or whether it was added at a later juncture. Thus, Pritchard prompts us to imagine Tempo who is fitted from birth with a highly reliable device, which records the ambient temperature. Moreover, Tempo has grown up in a society where it is completely natural for one to consult the temperature-recording device in order to form beliefs about the ambient temperature.

Pritchard claims that "interestingly, in a case like this it seems entirely unnecessary for Tempo to know that this is a reliable belief-forming process or what the source of the reliability before he gain knowledge via this process" (2010*b*, 146). I think this is correct. Forming beliefs in this way is a disposition for Tempo and, plausibly, there is CRC involved. Having always been fitted with the device Tempo has a practical understanding of how his actions will affect his temperature beliefs and vice versa. For instance, he has a practical understanding that when he goes closer to a heater or the fireplace the quick silver is supposed to rise and that when he moves away it will drop. Or, that if the temperature changes while he is sitting still then some warm object must be near by, or a window has been opened.[71] The temperature-

[70] That is, in being aware of the facts, Alvin has surely deemed his belief-forming process reliable and, as time goes by, it can start being one of his dispositions.

[71] Put another way, Tempo has acquired knowledge of the sensorimotor contingencies that accompany his continuous interaction with the device. For a full account of how sensorimotor knowledge is constitutive of perception see (Noë 2004). "The basic claim of the enactive approach is that the perceiver's ability to perceive is constituted (in part) by sensorimotor

recording device is not just a thermometer on the wall for Tempo. Instead, he is continuously interacting with the device as he moves around generating outputs which recycled as inputs drive his cognitive character along with respect to his temperature-movement related beliefs. Moreover, through all these interactions, it can be safely assumed that the device has been deemed reliable. Had the device told Tempo that it is cold while he is next to the fireplace, he would not have trusted it. Therefore, the device has been appropriately integrated within his cognitive character even though Tempo might not even be aware of its existence.

But then, Pritchard asks, what if the agent is fitted with the device at a later stage? So, imagine Tempo* who comes out of a comma with this device fitted, while being non-culpably unaware that this device has been artificially implanted in him (Pritchard 2010*b,* 148). Such a change, argues Pritchard, cries out for the agent to take a reflective stance on the epistemic standing of this change, and in its absence we cannot say that Tempo* can gain knowledge on the basis of his newly fitted device. Interestingly, however, Pritchard claims that as time goes by this intuition lessens. For if Tempo* has been fitted with the device, say for three years "there is now a track-record of beliefs formed via this process which have generally cohered with the beliefs formed via Tempo*'s cognitive abilities (and if they hadn't cohered, we may suppose, then Tempo* would have spotted this and responded accordingly)" (Pritchard 2010*b,* 148). That is, the new belief-forming process has now both become a disposition for Tempo* and has been deemed trustworthy. Plausibly, moreover, within these three years Tempo* has become able to reciprocally interact with the device such that it can count as his cognitive ability. In other words, given that Tempo* has been fitted with the device for a sufficiently long period of time, our intuitions on his ability to acquire knowledge on its basis become more supportive, as he may have plausibly satisfied the 3+1 criteria.

knowledge (i.e., by practical grasp of the way sensory stimulation varies as the perceiver moves)". (Noë 2004, 12). "What the perception is, however, is not a process in the brain, but a kind of skillful activity on the part of the animal as a whole". (Noë 2004, 2). "Perception is not something that happens to us or in us, it is something we do". (Noë 2004, 1). Sensorimotor dependencies are relations between movements or change and sensory stimulation. It is the practical knowledge of loops relating external objects and their properties with recurring patterns of change in sensory stimulation. These patterns of change may be caused by the moving subject, the moving object, the ambient environment (changes in illumination) and so on.

Now, before closing this section, one last remark is in order here. When discussing the Temp case where the thermometer is hanging on the wall, and Temp knows what the source of the reliability is, I claimed that it would be better analyzed in testimonial terms. So, one might fairly wonder: why not try to analyze any instance in which an agent is in cognitive contact with some epistemic artifact in testimonial terms? Although this might seem a promising strategy, it is not going to work for the same reason that HEMC cannot account for every case of an agent's deployment of an artifact. For, in many cases, the agent's true belief depends very deeply on his ongoing interaction with the epistemic artifact in such a way that no causal explanation of how the agent *formed* his true belief will be possible or available in linear/testimonial terms.

Consider, for instance, telescopic observation. Making observations through a telescope is a dynamic process that requires a great deal of experience in operating the artifact, and a great amount of background knowledge to understand what one is looking at. By moving around the telescope while adjusting the lenses, the agent delivers outputs (shapes on the lens of the telescope) which recycled as inputs drive the agent's cognitive character along, so as to, eventually, identify—that is, *see*—recognizable objects in space (e.g., stars, planets, comets, galaxies et cetera). The epistemic artifact actively drives the agent's cognitive mechanism in a continuous and highly interactive way. Therefore, agent and telescope should be considered as a coupled system, and the overall process as a case of extended cognition and not as merely a case of an agent using an instrument. That is, the interaction between the agent and the telescope is not linear such that the two systems can be neatly decomposed by having their function described in terms of distinct inputs and outputs from the one to the other. Accordingly, it would be a vain attempt to analyze one's knowledge of stellar facts in terms of (artificial) testimony, whereby the telescope provides the agent with fully articulated pieces of information, which she must accept or deny.

3.4) Extending the Cognitive Character

To recap, an epistemic agent's cognitive character—i.e., her cognitive system—may be extended beyond her belief-forming faculties that make up her cognitive agency by incorporating belief-forming processes, which rely for the most part on environmental elements. In order to count as genuine parts of the agent's cognitive character, these belief-forming processes must meet the 3+1 criteria. That is, they must be trustworthy dispositions in the sense of being reliably available, typically invoked and automatically endorsed. This also implies (though, of course, does not entail) that they are objectively reliable in the sense that they tend to produce true rather than false beliefs (otherwise, we may suppose, the processes wouldn't be typically invoked, nor would their results be automatically endorsed). Moreover, these dispositional, reliable, belief-forming processes must be appropriately integrated within the rest of the agent's cognitive character. One way to test whether these processes are so appropriately integrated is to check whether by employing them, the epistemic agent delivers outputs, which recycled as inputs drive his overall cognitive character along (i.e., CRC is manifested between the internal and the external elements of the process).

Therefore, the employment of hardware external elements such as epistemic artifacts, calculators, microscopes, telescopes, pen and paper can occasionally count as genuine extensions of one's cognitive character, depending on whether they meet the aforementioned criteria. It should be interesting, however, to see whether something analogous applies to what would count as a software extension.

3.4.1) Languages as Software Extensions

In our discussion in chapter 2, we briefly mentioned that from the HEC point of view, the development of language might have been, to a certain degree, the outcome of the human need to externalize their thoughts to the public space so that they can more easily manipulate them. Let us now take a closer look at this idea.

Drawing on Vygotsky's ideas as vindicated by recent bodies of developmental research (see (Berk & Carvin 1984)), Clark suspects that self-directed speech (be it vocal or silent inner rehearsal) is a crucial cognitive tool that allows us to highlight the most puzzling features of new situations, and to direct and control our own problem-solving actions (Clark 1998, 164). Of course, as he further notes, the effect of language on human thought need not be restricted to speech, since written language may have similar, and possibly more powerful, results: "in constructing an academic paper, for example, it is common practice to deploy multiple annotated texts, sets of notes and files, plans, lists and more. The writing process often depends heavily on manipulations of these props—new notes are created, old ones juxtaposed, source materials are wielded on and off of work surfaces etc" (Clark 1998).

Briefly, the main idea is that language in general, and words in particular, enable us to capture abstract ideas and rich experiences in memory. This has the direct effect of allowing thoughts to become objects of further attention and reflection, opening them up to a range of further mental operations. This *feedback of one's thoughts to one's own cognitive system* gives rise to the distinctively human capacity of meta-cognition, or, as Clark calls it, "second order cognitive dynamics" (1998, 177). This capacity to externalize one's thoughts in *recyclable* linguistic representations can be far more active and transformative than one may initially think, since the particular linguistic abilities one possesses may guide or restrain one's ongoing trains of thoughts. Take the construction of a poem for example: "we do not simply use the words to express thoughts. Rather, it is often the properties of the words (their structure and their cadence), which determine the thoughts that the poem comes to express. A similar partial reversal can occur during the construction of complex texts and arguments. By writing down our ideas we generate a trace in a format that opens up a range of new possibilities. We can then inspect and re-inspect the same ideas, coming at them from many different angles and in many different frames of mind" (Clark 1998, 176)

The moral, Clark claims, is that public language and text play more than just a preserving-and-communicating-ideas role; "instead, these external resources make available concepts, strategies and learning trajectories which are simply not available to individual un-augmented brains. Much of the true power of language lies in the underappreciated capacity to re-shape the computational spaces which confront intelligent agents" (Clark 1998). Indicatively, some of the distinctively transformative effects of language on our biological cognitive systems, as Clark enumerates them (1998, 169-173), are memory augmentation, attention and resource allocation, and data manipulation and representation.

Interestingly, in a somewhat similar vein, but drawing inspiration from complex systems and chaos theory, Logan (2003; 2006; 2008) presents the idea that speech is the first proper language embedded in an evolutionary series of languages, preceded by pre-verbal proto-languages (tool making, social intelligence, and mimetic communication) and followed by more task, or domain specific languages such as written language, mathematics and science. According to this picture, each new language emerged from the previous forms of language as a bifurcation to a new level of order in response to an information overload that the previous set of languages couldn't handle. What is strikingly similar to Clark's view is that Logan holds that a word packs a great deal of experiencing into a single utterance or sign. "A concept in the form of a word links many percepts of an individual and, hence, extends the brain's capacity to remember. Words as concepts are a form of *artificial memory* which creates *artificial connections*. Words bring order to a chaotic mind filled with memories of a myriad of experiences. Language is an emergent order" (Logan 2006, 153). We thus see, that just as Clark thought, so Logan holds that language serves *two* and not just one fundamental function; it is obviously a form of (i) communication, but it is also a form of (ii) information processing.

As a result, on the basis of those two authors, it is tempting to claim that the use of language is a trustworthy dispositional cognitive process that actively drives the cognitive loops of the agents who possess it along the lines of the CRC phenomenon. But is language, or more accurately, public language and text, the only software external artifacts that cognitive agents use so as to extend their cognitive characters?

3.4.2) Scientific Theories as Software Extensions

Interestingly, as we briefly mentioned before, Logan claims that

> speech, writing, math, science, and computing form an evolutionary chain of languages. Each of these activities can be considered as a separate language because each allows us to think differently, create new ideas and develop new forms of expression. Another consideration is that each of these five forms of language possesses its own unique semantics and syntax and hence qualifies as a language in itself according to criteria set by classical linguistics (2003, 3).

I will here concentrate on the discussion of the interesting case of scientific theories: if scientific theories do qualify as languages and thereby, as software artifacts, then according to HEC and COGA$_{weak}$, they could also count as belief-forming processes that extend the epistemic agent's cognitive character beyond his natural belief-forming faculties.

3.4.2.1) A Hint: Observations Are Theory-Laden

A hint that this is so comes from the old problem of theory-laden observations within the domain of philosophy of science. Briefly speaking, the validity of scientific theories depends on their accordance with empirical observations. It has been claimed, however, that observation involves perception as well as other underlying cognitive processes. As Kuhn says, "something like a paradigm is a prerequisite to perception itself. What a man sees depends both upon what he looks at and what his previous visual-conceptual experience has taught him to see".[72] That is, observations heavily depend on some underlying understanding (which stems from the already existing scientific theories and commonsensical habits of thought) of the way in which the world functions, and that understanding influences what is perceived, noticed, or deemed trustworthy of consideration. Therefore, the argument goes, since empirical observations presuppose a theoretical understanding, they cannot be the final arbiters of the validity of scientific theories.

Historically, the issue first emerged between Hempel (1966, 1970), who defended the distinction between observational and theoretical terms, and Hanson (1961; 1969) who maintained the theory-laden thesis of observation. Specifically, according to Hanson, not only are the observational sentences theory-laden but the observations themselves are theory-laden (1969):

> In short we usually "see" through spectacles made of our past experience, our knowledge, and tinted and molted by the logical forms of our special languages and notations. Seeing is what I shall call a "theory-laden" operation, about which I shall have increasingly more to say.

[72] Briefly speaking a paradigm refers to the scientific theory along with its auxiliary hypotheses that is most widely accepted by the scientific community and on the basis of which the latter conducts its research. Kuhn's focus was on the scientific progress as a whole, comprising of the individual scientists, paradigms, scientific communities and so on. I am here only interested in the relation between the individual scientists and the scientific theories they are working on.

Famously, the debate also has a counterpart in the philosophy of mind, as it was taken up by Churchland (1979; 1988; 1989) and Fodor (1984; 1988). Fodor, by appealing to illusions such as the Muller-Lyer experiment whereby the subjects' knowledge of the illusion does not alter their defective impressions, thinks that perceptual processes are modular (i.e., independent, closed, domain-specific processing modules). So, by definition, bodies of theory that are inaccessible to the modules do not affect the way the perceiver sees the world. Churchland, on the other hand, relying on studies such as those utilizing the ambiguous pictures of rabbit/duck and young/old woman, argues that higher cognitive processes can have an impact on visual processes. Specifically, higher order theories provide the agents with internal representations, which pick out important distinctions and structures in the external world. When the input to the agent's perceptual processes is variegated, or noisy, and thereby not clearly represented, these representations allow the agent to "respond to those inputs in a fashion that systematically reduces the error messages to a tickle. These I need hardly remind, are the functions typically ascribed to "theories"" (1989, 177).

Considerations such as those of Hanson and Churchland have widely been thought to produce a relativistic picture of science—and possibly epistemology as well—whose most prominent proponents are thought to be Feyerabend (1975) and Kuhn (1962) (the latter quite possibly unjustly though). As one of Kuhn's most infamous passages goes: "In so far as [the scientists'] recourse to that world is through what they see and do, we may want to say that after a revolution scientists respond to a different world" (1962, 110). Fortunately, however, modern cognitive psychology points away from relativism, at least as far as the theory-ladenness of observations is concerned, even though the phenomenon is not altogether denied.

In particular, Anna Estany (2001, 208) holds that

> The beliefs of the higher or more fundamental level influence how perceptual units are interpreted by the lower levels [...] Humans use both types of processes in perception because each have characteristic advantages and disadvantages. Thanks to top-down processes we can recognize patterns with incomplete or degraded information. Moreover, top-down processes make perception faster, but they can induce us to make mistakes in a perception by relying on previous knowledge.

However, even though our perceptual systems do get guidance from higher order expectations, it has been pointed out that when attention is caused by the mismatches between expectation and reality the inputs from the arousal system constitute a "reset wave" making it possible not to fall into arbitrary, relativistic errors of perception (Estany 2008, 213).

Similarly, Brewer and Lambert (2001), exploring the literature on relevant experiments, concede that "perception is determined by the interaction of top-down theory information and bottom-up sensory information" (178):

> However, note that in all of the above cases the stimuli were either ambiguous, degraded, or required a difficult perceptual judgment. In these cases the weak bottom-up information allowed the top-down influences to have a strong impact on perceptual experience. It seems likely that strong bottom-up information will override top-down information. [...] Thus, the top-down/bottom-up analysis allows one to have cases of theory-laden perception, but does not necessarily lead down the slippery slope of relativism.

For the purposes of the present chapter, however, the resolution of the problem of whether the theory-ladenness of observations may or may not lead to relativism is less important than its very existence. For it shows that scientific theories can be seen as cognitive dispositions whose employment actively drives the agents' cognitive character by creating outputs (observations) which recycled as inputs drive the agent's overall cognitive system along, resulting in further observations or scientific theories, which are hopefully—as Estany, Brewer and Lambert argue—still empirically testable. The theory-ladenness of the scientific process may sometimes lead to mistakes, but for the most part, when the input is ambiguous, noisy, or variegated it is an important facilitatory effect that boosts the scientists' performance in several ways. In particular, Brewer and Lambert note that background theories affect not only the scientists' perceptual processes but they also play a significant role in other aspects of the scientist's cognitive and epistemic life including attention, data evaluation and interpretation, data production, memory, and communication.[73]

[73] Cf. Clark's (1998, 169-173) "6 ways".

3.4.2.2) Extended Scientific Problem-Solving

The discussion of the previous section was only intended as a hint that scientific theories may be seen as software artifacts that extend one's cognitive character. Let me explain why. The main idea was that once one becomes fluent with a scientific theory, the way one perceives the world (in qualitative terms), the aspects of the world one attends to and considers worthy of consideration, and the way one stores and communicates one's experiences will be fundamentally altered. In other words, scientific theories have a very strong impact on one's point of view at the external world, allowing one to observationally interact with it in ways that would otherwise be unavailable, and this, in turn, affects back the ways one does science. Scientific theories, moreover, are external in the sense that no one is born with them inscribed on one's neural apparatus. Theories, instead, are acquired through a long period of training and practice during which scientists interact with teachers, professors, textbooks, scientific equipment, and so on. And once scientists become masters of such externally derived theories, the cognitive operations (including their observations) they are able to perform are qualitatively altered and significantly enhanced. Hence, scientific theories may be seen as external software epistemic artifacts that extend one's cognitive character. Here is the problem, however, and why the above can be nothing more than a hint. In order to resist this picture, the opponent of cognitive extension does not need to deny that once such external scientific theories are appropriately internalized they will have dramatic effects on the epistemic agent's cognitive loops. Crucially, however, he will further claim that all processing, including making theory-laden observations, will be exclusively performed within the scientist's head. Why, then, should this count as a case of cognitive extension?

Clark, of course, seems to be aware of this line of arguing against languages as software extensions, and this is why he does not restrict himself in claiming that all languages do is to facilitate or enhance one's inner processes of thought and reason. Instead, the examples Clark uses in order to illustrate his point involve agents who physically manipulate external linguistic symbols and representations so as to achieve cognitive tasks that would otherwise be infeasible. So are there any similar examples from the scientific domain, which could motivate the view that scientific theories can count as software cognitive artifacts in a similar way?

Let us move back to the example of long multiplication that was first presented in section 2.1. Drawing on McClelland et al. (1986), Giere and Moffatt (2003) claim that human brain networks have evolved for and are best at completing and recognizing patterns in input provided by the environment. But, the question is that "if something like that is correct, how does man do the kind of linear symbol processing required for activities such as using language and doing mathematics"? (Giere & Moffatt 2003, 302). "The answer given by McClelland et al.", they note, "was that man does the kind of cognitive processing required for these linear activities by creating and manipulating external representations. These latter tasks can be done by a complex pattern matcher" (*ibid.*). So, again, think about the process of solving the multiplication problem '987 times 789'.

> This process involves an external representation consisting of written symbols. These symbols are manipulated, literally by hand. The process involves eye-hand motor coordination and is not simply going on in the head of the person doing the multiplying. The person's contribution is: (1) setting up the problem in a physical form; (2) doing the correct manipulations in the right order; (3) supplying the products for any two integers, which can be done easily from memory (Giere & Moffatt 2003, 303).

The rest of the mathematical cognitive task, however, is, literally, performed externally on the basis of the interaction between brain, hand, pen, and paper. Here is then an example according to which mathematics—again, a language according to Logan's view—can be seen as a software artifact that extends an agent's epistemic cognitive character. But if this analysis is correct, then very similar analyses can be provided for the solution of scientific problems, which involve the physical manipulation of external scientific symbols and formulas.

Take, for example, the use of chemical formulas in organic chemistry as introduced by Berzelius and Dumas in the early nineteenth century.

> Assuming that the basic constituents in reactions are conserved, one can represent chemical reactions by equations in which the numbers of all constituents are the same on both sides of the equation. That is, the equation must balance. One can literally do theoretical chemistry by manipulating these symbols in the following example: (Giere & Moffatt 2003, 304)

$$CH_4 + 2O_2 = CO_2 + 2H_2O \quad \text{(The burning of methane)}$$

Such formulas are clearly external representations that form part of an extended cognitive system that allows scientists to explore possible reactions in organic chemistry. "That is, the cognitive process of balancing an equation does not take place solely in the head of some person, but consists of interactions between a person and physical, external representations" (*Ibid.*). So here is an example of a scientific theory that does not only alter the agent's inner cognitive processes, but also allows him to externalize his problems in symbols, whose physical manipulation enable him to come up with solutions that, arguably, would otherwise be unavailable. This example, however, is only one out of a vast number of similar cases and anyone who solved problems in mathematics, physics, chemistry, logic, or even biology at school could come up with one's own examples.

So, although more remains to be said on this matter, my tentative conclusion is that it could be the case that scientific theories, like public language and text, are a software external element that can extend one's cognitive character beyond one's cognitive agency—that is, beyond one's natural cognitive faculties. The fact that, as in the case of language so in the case of science too, we sometimes employ them without even realizing that we do so has nothing to do with their status of being external, cognitive, and consequently epistemic artifacts. Rather, it is an indication that they are so appropriately integrated within our cognitive characters that they are "transparent equipment": "equipment (like the carpenter's hammer) with which we are so familiar and fluent that we do not think about it in use, but rather rely on it to mediate our encounters with a still wider world. […] And it is this bundle of taken-for-granted skills, knowledge and abilities that—or so I am suggesting—quite properly structures and informs our sense of who we are, what we know, and what we can do" (Clark 2006, 106). They are the cognitive traits that make up our cognitive *characters*.

3.5) Conclusion

To recap, I have argued that COGA$_{weak}$ and HEC are complementary accounts. On one hand, COGA$_{weak}$ is an attempt to accommodate the ability intuition on knowledge (i.e., knowledge must be the product of cognitive ability). HEC, on the other hand, is an attempt to recognize a process as part of one's cognitive system despite whether it is wholly realized within an agent's head, or not.

Interestingly, the criteria set forth by COGA$_{weak}$ in order for a belief-forming process to count as knowledge-conducive (i.e., be one's cognitive ability) are the same as the criteria suggested by HEC in order for a process to be part of one's overall cognitive mechanism. I have been referring to them as the 3+1 criteria: 1) the process must be reliably available and typically invoked (i.e., it must be one of the agent's habitual cognitive routines), 2) it must be more-or-less automatically endorsed and not usually subject to critical scrutiny (i.e., it must be trustworthy/reliable), 3) it should also be easily accessible as and when required (i.e., it must have been appropriately integrated within the agent's overall cognitive mechanism/character), and +1) there must be a continuous reciprocal interaction between the target process and the agent's natural cognitive faculties.

Moreover, this agreement on the fundamental tenets of the two views provides us with a principled account of the way in which epistemic agents may extend their knowledge-conducive cognitive characters beyond their natural cognitive faculties by incorporating epistemic artifacts. Thus, COGA$_{weak}$ constitutes a formulation of the ability intuition on knowledge, which also bears the important advantage of allowing the acquisition of knowledge on the basis of epistemic artifacts.

Finally, having checked the 3+1 criteria against a series of thought experiments which are meant to unravel our epistemological intuitions and intuitions related to the nature of the mind, we examined the claim that it is not only hardware artifacts that may extend our cognitive characters but it may well be the case that our minds can be extended via the employment of software tools such as languages. Even though more remains to be said on this matter, we further noted that if scientific theories can be seen as languages then, arguably, they too could be considered as software artifacts that extend one's cognitive system by actively driving (and restraining) one's ongoing cognitive loops. As Lakatos once wrote: "[we, scientists,] *use our most successful theories as extensions of our senses*" (1970, 107, emphasis in the original).

PART 2

CHAPTER 4

Weak Epistemic Anti-Individualism

4.1) Introduction

In the previous three chapters, we focused on the importance of the ability intuition on knowledge, namely the idea that knowledge must be the product of one's cognitive ability. Through the consideration of several thought experiments, Chapter 1 was mainly dedicated to the acknowledgement of the necessity of the ability intuition for any adequate account of knowledge. Another important consideration that emerged through the discussion of the first chapter was the realization that the ability intuition on knowledge, as captured by virtue reliabilism, cannot account for the acquisition of knowledge on the basis of epistemic artifacts. We thus introduced a necessary virtue reliabilistic condition on knowledge, which has recently been put forward by Pritchard (2010*b*), namely COGA$_{weak}$. COGA$_{weak}$ appears to avoid many of the problems that trouble virtue reliabilism, while it can also reconcile the ability intuition on knowledge with the fact that we so often gain knowledge on the basis of the operation of epistemic artifacts. In particular, according to COGA$_{weak}$, so long as one's true belief is significantly creditable to one's cognitive agency (i.e., one's organismic faculties of the brain/CNS), then the process by which one came to form one's true belief can count as *bona fide* cognitive ability. But, even though considering the operation of epistemic artifacts as cognitive abilities is epistemologically motivated, this is a radical claim that, in the absence of any metaphysical support, sounds rather implausible. Accordingly, Chapter 2 was devoted to the production of just this metaphysical support for this aspect of COGA$_{weak}$, through the consideration of the hypothesis of extended cognition (HEC). Finally, in Chapter 3, we tried to provide substantial evidence for the theoretical affinity between COGA$_{weak}$ and HEC. In particular, we argued that the criteria to be met by an external process such that it can count as part of one's cognitive system are the same as the criteria that must be satisfied so that a belief-forming process counts as knowledge-conducive. Given that both of these theories are meant to capture the notion of *cognitive ability*, this is exactly the outcome one would expect to find. Moreover, the fact that those two theories have been developed separately within different philosophical domains does not only render them mutually supportive and informative, but it is also a hint to their correctness.

With these considerations in mind, in this fourth chapter, I aim to explore the epistemological ramifications of understanding COGA$_{\text{weak}}$ along the lines suggested by HEC. In particular, the fact that knowledge-conducive belief-forming cognitive processes are no more restricted within the bodily boundaries of the individual epistemic agents is a good indication of the exogenous, social nature of knowledge. Notice, however, that the claim is not yet going to be that knowledge might, in certain cases, be entirely social (i.e., I won't yet support *robust anti-individualism* in epistemology). This will be the topic of the last two chapters. Rather, I will here focus on the recognition of the *dual nature of knowledge*: the idea that, in most cases, *knowledge is essentially both social and individual.*

The starting point for this idea can be traced in the common element to be found in any virtue reliabilistic condition on knowledge. Below are the formulations of virtue reliabilism and COGA$_{\text{weak}}$ as we have encountered them in the previous chapters:

Virtue Reliabilism

- *S* knows that *p* if and only if *S's* reliable cognitive character is the most important necessary part of the total set of causal factors that give rise to *S's* believing the truth regarding *p*.

Or,

- *S* knows that *p* if and only if *S's* cognitive success is primarily creditable to *S's* cognitive character.

(The above two formulations are supposed to be equivalent, since Greco (2004) holds that credit attributions are tantamount, or, at least, very much akin to causal explanations).[74]

[74] As I will later argue, however, credit attributions differ to causal explanations in at least one significant respect. That is, credit attributions pick out only intentional agents, whereas causal explanations may refer to both intentional and non-intentional aspects of the world.

COGA $_{weak}$

- If *S* knows that *p*, then *S*'s true belief that *p* is the product of a reliable belief-forming process, which is appropriately integrated within *S*'s cognitive character such that her cognitive success is to a significant degree creditable to her cognitive agency.

Even though these two formulations of the ability intuition on knowledge are not the same—the most obvious difference being that virtue reliabilism is meant to be both a necessary and sufficient condition on knowledge—they nevertheless share a core insight of what knowledge is supposed to be. In particular, both accounts attempt to understand knowledge in terms of credit attributions. This commonality between the two accounts, however, should not come as a surprise. In trying to accommodate the ability intuition on knowledge, both views share another common idea: credit is usually attributable in cases of success through ability.[75] Accordingly, as Greco notes, knowledge—or, at least, a necessary aspect of it—is *creditable true belief* (2007, 57).

Notice a difference in the detail, however. Whereas virtue reliabilism requires that one's cognitive success be *primarily* creditable to one's *cognitive character*, COGA$_{\text{weak}}$ demands that one's true belief be merely *significantly* creditable to one's *cognitive agency*. As we noted in chapter 1, COGA$_{\text{weak}}$'s lenient demands with respect to the degree of creditability of one's cognitive success to one's cognitive agency not only helps COGA$_{\text{weak}}$ to account for cases of testimonial knowledge,[76] but focusing on one's cognitive agency—as opposed to one's cognitive character—allows COGA$_{\text{weak}}$ to also explain how it is possible for an epistemic artifact to count as part of one's cognitive

[75] Let me note, however, a subtle difference between the two proposals again. While Greco presents knowledge as true belief which is 'of credit', Pritchard insists on thinking about knowledge merely as 'creditable' true belief. These two notions are not the same. "For example, one's cognitive success could be creditable to one's cognitive agency without being at all of credit to one (perhaps the cognitive success is the result of an inquiry that one ought not to be pursuing, because, say, there are epistemically more desirable inquiries that one should be focusing instead" (Pritchard *forthcoming*, en. 26). While this distinction is not important to the present discussion, it is of great significance with respect to the debate on the value of knowledge. If, as Greco claims, knowledge is true belief, which is 'of credit', this is because knowledge is an achievement. Since achievements are finally valuable, knowledge turns out to be finally valuable, as well. However, considering cases such as the one mentioned above, or mundane instances of knowledge such as perceptual beliefs, Pritchard claims that knowledge is not always an achievement and so not finally valuable either. For further discussion on this issue, see (Pritchard 2010*a*, §2.4).

[76] We will return to testimonial knowledge shortly, in the next section.

character, such that one can gain knowledge on its basis. In such cases, even though one's cognitive success is primarily the product of the operation of the epistemic artifact, one's cognitive agency deserves significant credit, for it is one's cognitive agency that is first and foremost responsible for *integrating* into one's cognitive character the epistemic artifact, whose operation allows one to enjoy cognitive success.

Now, understanding knowledge in terms of credit attributions, as suggested in the lines above, will help us reveal the social aspects of individual knowledge. For as it will become apparent in the following sections, a preponderance of the individualistic true beliefs that amount to knowledge cannot be wholly creditable to the individual subjects whose knowledge status is each time assessed. I will begin with considerations that pertain to testimonial knowledge and I will then move on to examine cases where one's true believing is the product of epistemic artifacts. In all these cases, the cognitive success may well be significantly creditable to the cognitive agency of the individual subject whose knowledge status is being assessed. But the rest of the credit should, or so I will argue, be attributed to other specific individuals, or to the epistemic society in which the individual subjects are embedded. Finally, I will examine Hardwig's claim that this epistemic dependence on one's epistemic circle leads to either of the two following unpalatable conclusions: either scientific research and scholarship do not constitute knowledge at all, or such knowledge is not possessed by any individual alone, but by the epistemic community as a whole. Fortunately, there is a third conclusion to be drawn, originating from (but not jumping straight out of) externalist epistemology, and which points to the dual nature of knowledge.

4.2) Testimonial Knowledge

4.2.1) Introduction to Testimonial Knowledge

Testimonial knowledge has always been a central topic in epistemology. The reason is obvious: very many of our everyday beliefs appear to have testimonial origins. Accordingly, an adequate account of knowledge should be able to accommodate this powerful source of knowledge. Traditionally, the two opposing sides within the debate concerning testimonial knowledge are those of reductionism—which assigns the entire epistemic burden to the hearer—and non-reductionism—which shifts the epistemic burden to the speaker. Recently, however, Jennifer Lackey (2008) has put forward a dualist account of testimonial knowledge that appears to accommodate both of these seemingly opposing views. The reason, she claims, is that "it takes two to tango"; "an adequate view of testimonial justification or warrant needs to recognize that the justification or warrant of a hearer's belief has *dual* sources, being grounded in both the reliability of the speaker and the rationality of the hearer's reasons for belief" (Lackey 2008, 177). The aim of this section is to present the debate between reductionism and non-reductionism, introduce Lackey's view, and then check whether her dualist account is in line with $\text{COGA}_{\text{weak}}$. If this turns out to be the case, then $\text{COGA}_{\text{weak}}$ is well equipped to share Lackey's insight with respect to the epistemic burden distribution that testimonial knowledge points to, thus providing us with a first sense in which individual knowledge can at the same time be social in nature.

To start with, consider reductionism about testimonial knowledge, according to which the epistemic status of testimony is ultimately reducible to sense perception, memory, and inductive inference. As Hume (1977, 75)—who is often regarded (quite possibly unjustly) as the best-known supporter of reductionism regarding the epistemology of testimony—notes, "the reason why we place any credit in witnesses and historians, is not derived from any *connexion*, which we perceive *a priori*, between testimony and reality, but because we are accustomed to find a conformity between them."[77]

More precisely, reductionists ascribe to the *'positive reasons'* thesis, according to which justification or warrant is attached to testimonial beliefs only by the presence of appropriate positive reasons on the part of the hearers, thereby assigning *all* of the epistemic burden on the hearers' shoulders. Since these reasons cannot be testimonial (otherwise there would be circularity) they must depend on other epistemic sources that typically include sense perception, memory and inductive inference. Testimonial

[77] In a similar vein, Faulkner (2000, 587-8) claims that "it is doxastically irresponsible to accept testimony without some background belief in the testimony's credibility or truth", and "an audience is justified in forming a testimonial belief if and only if he is justified in accepting the speaker's testimony." Or, consider Fricker (1994, 149-50): "the hearer should be discriminating in her attitude to the speaker, in that she should be continually evaluating him for trustworthiness throughout their exchange, in the light of the evidence, or cues, available to her".

justification, or warrant is, therefore, ultimately reducible to the justification/warrant of these basic epistemic sources. Having these considerations in mind, Lackey formulates reductionism thusly:

> *Reductionism*
> For every speaker, A, and hearer, B, B believes that *p* with justification/warrant on the basis of A's testimony if and only if:
> (R1) B believes that p on the basis of A's testimony;
> (R2) B has sufficiently good non-testimonial positive reasons to accept A's testimony. (Lackey 2008, 145)

As Lackey notes, however, the possession of appropriately positive reasons does not necessarily guarantee to the hearer the reliability of the speaker's testimony.[78] Consider, for example, Max who has known Ethel for the last ten years, over the course of which, he has acquired excellent positive grounds for thinking that Ethel is a reliable source of testimony. Currently, however, Ethel is going through a personal crisis that no one knows about, and, so, in a state of distress, reports to Max that her purse has been stolen, even though she has no reason to think that this is the case. Ironically enough, however, and unbeknownst to Ethel, it turns out that her purse was in fact stolen when she was at the coffee shop, earlier that same day.

What this Gettier-case demonstrates, Lackey (2008, 152) explains, is that despite the fact that the hearer has excellent positive reasons for accepting the speaker's testimony, the speaker acts "completely out of epistemic character", delivering an unreliable report, which though it turns out to be true, prevents the hearer from acquiring knowledge. Therefore, as mentioned before, the possession of appropriately positive reasons does not necessarily guarantee to the hearer the reliability of the speaker's testimony and so, Lackey concludes, reductionism is not an adequate account of testimonial knowledge.

So, let us turn to non-reductionism, according to which testimony is just as epistemically basic as sense perception, memory, and inductive inference. Such a view can be traced back to the work of Reid (1983, 281-2) according to which "the wise author of nature hath planted in the human mind a propensity to rely upon human testimony before we can get a reason

[78] Lackey supports her claim through the consideration of two examples, namely 'Nested Speaker' and 'Unnested Speaker' (Lackey 2008, 148; 152).

for doing so". So, on the non-reductionist view, acquiring testimonial knowledge does not require the possession of any positive reasons on the part of the hearer; instead, as Burge (1993, 467) explains, "a person is entitled to accept as true something that is presented as true and is intelligible to him, unless there are stronger reasons not to do so". Or, consider Weiner (2003, 257) who, in a similar vein, holds that "we are justified in accepting anything that we are told unless there is positive evidence against doing so".[79]

Notice the commonplace in all the aforementioned views; while the absence of any negative reasons *is* necessary for the acquisition of testimonially based knowledge, the presence of positive reasons is not. Put another way, non-reductionists hold that so long as there are no relevant undefeated defeaters,[80] hearers can acquire testimonially based knowledge merely on the basis of a speaker's testimony, thereby seemingly shifting the entire epistemic burden from the hearer to the speaker.[81]

Accordingly, Lackey formulates her version of non-reductionism thusly:

> *Non-Reductionism*
> For every speaker, A, and hearer, B, B knows that *p* on the basis of A's testimony if and only if:
>
> (NR1) B believes that p on the basis of the content of A's testimony;
> (NR2) B has no undefeated (psychological or normative) defeaters for A's testimony;
> (NR3) It is true that *p*. (Lackey 2008, 158)

Lackey, however, goes on to test non-reductionism against the 'Incompetent Agent' (158) where an unreliable speaker testifies to a hearer. But, the hearer possesses no relevant undefeated defeaters and so, according to non-

[79] In a similar spirit, Audi (1998, 142) claims that "gaining testimonially grounded knowledge normally requires only having no reason for doubt about the credibility of the attester."

[80] It is here important to introduce the two relevant types of defeaters that could affect one's acquisition of testimonial knowledge. First, there are psychological defeaters, which are beliefs or doubts that are had by the hearer and which indicate that the hearer's beliefs are either false or unreliably formed. Notice that psychological defeaters may not be objectively correct. Second, there are normative defeaters, which are doubts or beliefs that the hearer ought to have, and which indicate that the hearer's beliefs are either false or unreliably formed. In other words, normative defeaters are beliefs or doubts that the hearer should have (despite whether or not the hearer does actually have them), given the presence of certain available evidence.

[81] I here say 'seemingly' because, as it will become apparent later on, to possess no undefeated defeaters against a testimonial report is actually a condition that requires a fairly active epistemic stance on the part of the hearer.

reductionism, must ultimately accept the proffered defective statement as true. Accordingly, just as in the case of merely possessing positive reasons, the mere absence of negative ones does not guarantee to the hearer the reliability of the speaker's testimony and so, Lackey suggests, non-reductionism is in need of the further condition (NR4):

> (NR4) The speaker's testimony is reliable or otherwise truth-conducive.

Still, however, Lackey argues, if the receiver of testimony is either insensitive to the relevant undefeated defeaters—even though they are evidentially present to him—or oversensitive to them, then he will be unjustified/unwarranted in accepting the speaker's testimony.[82] Accordingly, we must ensure that the hearer in question has the capacity for and is appropriately sensitive to the relevant defeaters. Hence, non-reductionism must again be supplemented by the further condition (NR5):

> (NR5) The hearer is a reliable or properly functioning recipient of testimony. (Lackey 2008, 164)

Yet again, all of the five conditions that have been so far proposed appear inadequate. The reason is that the counterexample of the "Insular Community" (Lackey 2008, 164-5)—in which the hero happens to ask for directions the only reliable testifier in a city whose members always deceive the 'outsiders'—accentuates the need that the environment wherein testimony is exchanged must be suitable for the reception of reliable testimony. Consequently, non-reductionism must be strengthened with one last condition:

> (NR6) The environment in which B receives A's testimony is suitable for the reception of reliable testimony. (Lackey 2008, 167)

Finally, having formed non-reductionism along the lines of the above six conditions, Lackey goes on to test it against one last counterexample, namely 'Alien' (Lackey 2008, 168-9): Walking in the forest, Sam sees someone who looks like an alien dropping a book. Sam recovers the book and notices that it is written in what appears to be English and it looks like to what we on

[82] Lackey refers to these two examples as 'Good-Natured' and 'Compulsively Paranoid'. See (Lackey 2008 160; 161).

Earth would call a diary. By reading the first sentence of the book, Sam forms the corresponding belief that tigers have eaten some of the inhabitants of the author's planet. In reality, the book is a diary written in English and it is true and reliably written in it that tigers have eaten the aliens. Sam is also a properly functioning recipient of testimony and he is situated in an environment that is suitable for the reception of reliable reports.

Now, despite the fact that all the above six conditions are satisfied, it seems implausible to accept that Sam gains knowledge in this case. The reason, Lackey explains, is that Sam holds no positive reasons on behalf of the speaker's testimony; he knows nothing about aliens, he has no beliefs about their reliability as testifiers, he has no idea about the purpose of alien 'diaries', he has no common-sense alien-psychological theory, he has no beliefs about the reliability of the author of this book and so on. In the absence of such positive reasons, Lackey suggests, the only rational choice for Sam is to withhold belief.

Overall, then, what the Alien counterexample purports to demonstrate is that in the absence of any positive reasons for the reliability of the speaker's testimony, it is *not rational* for the hearer to accept the target testimony. Therefore, while reductionism as expressed through the positive reasons thesis is not a sufficient account of testimonial knowledge, it nevertheless seems to capture a necessary aspect of it. Therefore, Lackey proposes, we are in need of a dual account of testimonial knowledge that will involve both reductionist and non-reductionist conditions and which she formulates as follows:

> *Lackey's Dualism*
> For every speaker, A, and hearer, B, B knows (believes with justification/warrant) that *p* on the basis of A's testimony only if:
> (D1) B believes that *p* on the basis of the content of A's testimony;
> (D2) A's testimony is reliable or otherwise truth conducive;
> (D3) B is a reliable or properly working recipient of testimony;
> (D4) The environment in which B receives A's testimony is suitable for the reception of reliable testimony;
> (D5) B has no undefeated (psychological or normative) defeaters for A's testimony;
> (D6) B has appropriate positive reasons for accepting A's testimony. (Lackey 2008, 177-8)

To properly appreciate the motivating idea of Lackey's dualist account, remember that reductionism puts all of the epistemic responsibility on the hearer while non-reductionism assigns the entire epistemic work to the speaker. On the contrary, the main idea that motivates Lackey's dualist account is the realization that the epistemic burden must be distributed across both the speaker and the hearer. As Lackey vividly puts it, "it takes two to tango", because "an adequate view of testimonial justification or warrant needs to recognize that the justification or warrant of a hearer's belief has *dual* sources, being grounded in both the reliability of the speaker and the rationality of the hearer's reasons for belief" (Lackey 2008, 177).

In addition, there are two interrelated clarifications that are in order here. First, notice that since the epistemic burden of testimonial justification/warrant is distributed across both the speaker and the hearer, the demand for the acquisition of positive reasons on the part of the hearer is not as strong as reductionists would require it to be. That is to say, the 'positive reasons' condition (D6) is not meant as a sufficient condition for acquiring testimonial knowledge. Rather, positive reasons are only required in order to render the hearer's acceptance of the speaker's testimony "rational, or at least, not irrational" (Lackey 2008, 180).

And second, notice that even though it has been argued that having positive reasons for accepting one's testimony requires from the hearer to have all kinds of knowledge about people, their areas of expertise and their psychological propensities, which knowledge most subjects lack, this is not an actual problem for dualism. Granted, to accept one's testimony as true on merely positive reasons requires a great deal of relevant knowledge, which is implausible to assume that normal subjects have. But, requiring positive reasons that can make my acceptance of one's testimony rational, or at least not irrational is a far less demanding requisite that can be satisfied in much simpler ways. In particular, Lackey provides three types of inductively based positive reasons that could allow normal subjects to identify reliable (or unreliable) testimony.[83]

[83] Although the following types are originally meant for the provision of positive reasons for accepting one's testimony, it is true that they can also be used equally well for the seemingly diametrically opposite process of coming up with undefeated defeaters for rejecting one's testimony.

The first type includes criteria for individuating epistemically reliable contexts and contextual features:

> Specifically, even if B has not observed a general conformity of reports delivered in contexts of kind C and the truth, B may have observed the general conformity of reports delivered in contexts of kind C and the truth. So, if B believes that A's report is delivered in a C-context, then this, combined with B's inductive evidence regarding contexts of kind C, may give B an epistemically relevant positive reason for A's testimony. (Lackey 2008, 182)

For example, one may more easily accept the reports proffered in an astronomy lecture or found in *National Geographic* than the reports made in an astrology lecture or found in the *National Enquirer*. Or, in a similar vein, one may more easily accept the report of a calm and coherent witness testifying a robbery a few blocks away than would accept the report of a confused person who smells of alcohol.[84] "Similar remarks can be made about countless other contextual factors such as facial expressions, eye contact, mannerisms, narrative voice and so on" (Lackey 2008, 182).

The second class of reasons pertains to criteria that can help us make distinctions between reports:

> In particular, even if B has not observed a general conformity between A's reports and the truth, B may have observed the general conformity of reports of kind R and the truth. Thus, if B believes that A's report is an instance of kind R, this, combined with B's inductive evidence regarding R-reports, may give B an epistemically relevant positive reason for A's testimony. (Lackey 2008, 182)

For instance, one may uncritically accept one's testimony of the time of the day, one's name, what one ate for breakfast, while one may adopt a more critical stance towards one reporting about political matters, the achievements of one's children, one's sexual performance, UFO sightings, and so on.

Finally, the third kind of criteria that Lackey puts forward are meant to help the hearer to distinguish between epistemically reliable and unreliable speakers:

[84] In relation to the previous footnote, see how this second case, as Lackey herself also suggests, is best explained in terms of either the possession or absence of undefeated defeaters, rather than the presence of positive reasons (2008, 181).

> Specifically, even if B has not observed the general conformity between A's reports and the truth, B may have observed the general conformity of speakers of kind S and the truth. Thus, if B believes that A is an S-speaker, then this combined with B's inductive evidence regarding S-speakers, may give B an epistemically positive reason for A's testimony. (Lackey 2008, 183)

Consider, for example, that when one tries to find one's way to a desired destination in an unfamiliar city, one may accept in a less hesitant manner the testimony of someone who seems to be a local passer-by than would accept the word of someone who looks like a tourist.

The upshot of the above considerations is that despite the arguments that point to the opposite direction, there is indeed a plethora of ways in which a hearer can render her acceptance of a speaker's testimony rational or, at least, not irrational, in the way that the 'positive reasons' condition of dualism demands.

In summary then, and before turning to COGA$_{weak}$, since the process of acquiring information through testimony so as to form the corresponding beliefs is an interactive exchange between the hearer and the speaker, the relevant beliefs can only become justified/warranted by conditions that pertain to both parties of the said exchange. This realization has made Lackey go beyond the debate between reductionism and non-reductionism, thereby wedding these two views in a single dual account.

4.2.2) Dualism in the Epistemology of Testimony and COGA$_{weak}$

Now, to see how virtue reliabilism fares with respect to testimonial knowledge, recall the following counterexample that we first encountered in chapter 1.

> *Jenny*[85]
> Jenny gets off the train in an unfamiliar city and asks the first person that she meets for directions. The person that she asks is indeed knowledgeable about the area, and gives her directions. Jenny believes what she is told and goes on her way to her intended destination.

[85] (Pritchard 2009, 68). It is adapted from Lackey's 'Morris case' (see Lackey 2007, 352)

Unless we want to deny a great amount of knowledge that we suppose we have, we must admit that Jenny gains knowledge in this way. Lackey (2007) has argued, however, that given the way Jenny gains knowledge, her cognitive character, it seems, has nothing to do with the true-status of her belief. Instead, it is the informant's cognitive character that is the most salient factor in the causal explanation of why Jenny believes the truth. So, according to virtue reliabilism, Jenny lacks knowledge that she in fact possesses.

As Pritchard explains, however, if we are indeed inclined to accept that Jenny gains knowledge in this case, this is because we are reading the example both in such a way that Jenny's cognitive character is to some significant extent—thought, of course, not primarily—involved in how she gets things right, and in a way that her environment is friendly for the reception of reliable testimony. In particular, Pritchard explains, to say that Jenny gains knowledge in this way, we must read the example in such a way that Jenny is in an epistemically friendly environment—i.e., the city that Jenny visits had better not be renowned for its dishonest informants. Were that to be the case, we would not credit Jenny with knowledge. Second, we presuppose some inclinations about Jenny's cognitive character. We expect that Jenny can distinguish between potentially reliable and clearly unreliable informants; we do not expect that Jenny would be happy to ask just anybody. For example, we anticipate that she would not ask someone who clearly looked like a tourist (i.e., an unreliable informant). "Had the first person she met been obviously mad, or a stereotypical tourist, for example, then we would expect her to move on to the next prospective informant down the street" (Pritchard *forthcoming,* 18). Moreover, we expect that she is able to distinguish between potentially reliable and clearly unreliable information and thereby that she would not believe whatever she was told, had it been obviously false (for instance to go past the city hall whereas, in fact, she is in a village). "Furthermore, if the manner in which the informant passed on the directions was clearly questionable—if the informant was vague, shifty, hostile, and evasive, say—then we would expect our hero to exercise due caution" (Pritchard *forthcoming,* 18). Had Jenny not been responsive to these epistemically relevant factors, we would not have normally attributed her with knowledge. We, therefore, see that it is not that Jenny's cognitive character has nothing to do with her believing the truth; it is just that the informant's role is more important.

Obviously, then, $\text{COGA}_{\text{weak}}$ can easily handle the Jenny case; although the cognitive success is not *primarily* creditable to Jenny—but to the stranger—Jenny, in so being responsive to the epistemically relevant factors, has the right sort of abilities and employs them in the right sort of way so as to accept the stranger's information such that her cognitive success is *significantly* creditable to her cognitive agency.[86] According to $\text{COGA}_{\text{weak}}$, therefore, Jenny can gain knowledge in this way.

So, in contrast to virtue reliabilism, $\text{COGA}_{\text{weak}}$ appears to satisfactorily handle testimonial knowledge. Should we also be confident that it is in agreement with Lackey's dualist account? And, more importantly to the purpose of this chapter, can it accommodate the epistemic burden distribution that Lackey's account brings to light? To answer to these two questions I will first consider condition D4, then I will move on to D6 and D5, D3 and I will finish with D2.

So, to begin with, consider condition D4 according to which *the environment in which B receives A's testimony must be suitable for the reception of reliable testimony*. The reason for which I take up this point first is that there is no straight reference to it in the formulation of $\text{COGA}_{\text{weak}}$. Nevertheless, notice that Pritchard's investigation of what goes wrong in the Jenny case begins with the point that it is important to note that to say that Jenny gains knowledge in that way, we must read the example in such a way that Jenny is in an epistemically friendly environment; it is not as if the city that Jenny visits is renowned for its dishonest informants. Were that to be the case, then we would not attribute knowledge to Jenny. So we see that the problem posited by the knowledge-undermining luck that attaches to a true belief when formed in an inappropriate environment goes far beyond unnoticed in the process of formulating $\text{COGA}_{\text{weak}}$.[87].And since $\text{COGA}_{\text{weak}}$ is only a

[86] Notice that integrating information acquired by external sources within one's cognitive character is itself a belief-forming process, which is reducible to more basic inductively and memory based belief-forming processes. Nevertheless, this kind of belief-forming process seems to be critical even though it is usually a transparent one.

[87] In fact, Pritchard recognizes the problem posited by the knowledge-undermining luck to be a central one. Accordingly, he elsewhere formulates a complete account of knowledge by combining $\text{COGA}_{\text{weak}}$ with an anti-luck condition on knowledge, namely the safety principle. Consider for example *Anti-Luck Virtue Epistemology: S knows that p if and only if S's safe belief that p is the product of her relevant cognitive abilities (such that her safe cognitive success is to a*

necessary condition on knowledge—and should this become a pressing point—there is nothing preventing us from adding to it a supplementary clause that could rule out the lucky acquisition of true beliefs due to the environmental inappropriateness.

Let us now turn to conditions D5 and D6 according to which *the hearer must have no undefeated defeaters against the speaker's testimony* and *the hearer must have appropriate positive reasons for the speaker's testimony*, respectively. The incentive for discussing these two conditions together is that they are jointly meant to ensure the rational or, at least, not irrational, acceptance of the speaker's testimony. Moreover, notice that the absence of any undefeated defeaters against, and the possession of positive reasons for a testimonial report could be thought of as the two sides of the same coin. To see why, pay attention to the fact that both conditions require that one is aware of, able and supposed to detect any such reasons, should they become evidentially available. The only difference is that in order to acquire testimonial knowledge, in the end, no undefeated defeaters must remain while positive reasons must have been acquired. Importantly, however, both conditions require an active stance on the part of the hearer in the sense that she must be in a continuous lookout for satisfying them. I shall return to this point later on in the discussion of condition D3.

Meanwhile, we can return to COGA$_{weak}$ to see how conditions D5 and D6 can be seen through the lens of this account. First, notice that Pritchard clearly acknowledges that to say that Jenny gains knowledge in this way, we presuppose some natural inclinations about her cognitive character. We expect that Jenny can distinguish between potentially reliable and clearly unreliable informants; we do not expect that Jenny would be happy to ask just anybody. For example we anticipate that she would not ask someone who clearly looked like a tourist (i.e., an unreliable informant), or that she would not trust an informant that is vague, hostile or evasive. Moreover, we expect that she is able to distinguish between potentially reliable and clearly unreliable information and that she would therefore not believe whatever she was told, had it been obviously false. Had Jenny not been responsive to such

significant degree creditable to her cognitive agency (Pritchard, *forthcoming*). And again, in (Pritchard *2010a*, 76) we can read: "knowledge is safe belief that arises out of the reliable cognitive traits that make up one's cognitive character, such that one's cognitive success is to a significant degree creditable to one's cognitive character".

epistemically relevant factors then we would not have normally attributed her with knowledge. Interestingly, notice that Jenny's responsiveness to these epistemically relevant factors can also be described in terms of the three types of inductively based positive and negative reasons that Lackey grants to epistemic agents for identifying reliable (or unreliable) instances of testimony; namely, (i) criteria for individuating epistemically reliable contexts and contextual features, (ii) criteria for distinguishing between reliable and unreliable reports and (iii) criteria for identifying epistemically reliable speakers.[88]

Therefore, we see that Jenny, in so being responsive to such epistemologically relevant factors, has the right sort of inductively based belief-forming processes and employs them in the right sort of way so as to appropriately integrate the information conveyed by the communicable content of the speaker's act of communication within the rest of her cognitive character. What is of further import, however, is to notice that since one's cognitive character has been described as consisting of one's perceptual cognitive faculties, acquired habits of thought, but also of one's memories and the entire doxastic system, we can see that to say that Jenny, in so being responsive to the epistemically relevant factors, appropriately integrates the speaker's information within her cognitive character is on a par with saying that Jenny renders rational or, at least, not irrational the acceptance of the speaker's testimony. Because to rationally accept a piece of information is to say that this information does not conflict with the rest of one's beliefs, or that the process of acquiring it does not conflict with the rest of one's doxastic attitudes. And this is exactly what Lackey's conditions D5 and D6 were intended for.

Let us now move on to the last condition that pertains to the hearer. D3 demands that *the hearer is a reliable or properly functioning recipient of testimony.* As we have seen, the reason for which Lackey includes this condition is to rule out cases in which the recipient of testimony either has positive or

[88] Notice that, as Lackey herself admits, this list is not meant to be exhaustive as there could be further inductively based ways to distinguish between the reliability and unreliability of testimonial reports (2008, 181). Nevertheless, the identification of reliable reports should not be thought of as being exclusively based on inductive reasons, as it may often be the outcome of reasons that have to do with the agent's memory; consider, for example, an agent assessing the coherence of information provided by a proffered report with the rest of his/her doxastic system.

negative reasons evidentially available to him but fails to properly appreciate them, or is oversensitive to them, thereby being viciously justified in accepting the speaker's testimony. What must be further noticed, though, is that Lackey makes condition D3 subtler by adding the qualification that the hearer must be a reliable or properly functioning recipient of testimony in a substantial way. In particular, in her defense of her dualist view against the Infant/Child Objection, Lackey argues (2008, ch.7) that the only meaningful way for the 'no undefeated defeaters' condition (D5) to be satisfied is substantively, as opposed to trivially. To better understand this point, she prompts us to consider the following: "if we impose a no-φing condition on X then there is a crucial difference between what we might call *trivial satisfaction* and *substantive satisfaction* of such condition, a difference that depends on X's capacity to φ. In particular, let us put forth the following:

> *Trivial Satisfaction:* If X does not φ merely because X does not have the capacity to φ, then X has trivially satisfied the no-φing condition.
> *Substantial Satisfaction:* If X has the capacity to φ and does not φ, then X has substantively satisfied the no-φing condition." (Lackey 2008, 198)[89]

Having this crucial distinction in mind, Lackey goes on to explain that if φ is an epistemological or moral condition, then only in the second case is X epistemologically or morally praiseworthy for satisfying it (Lackey 2008, 198). Therefore, conditions D5 and D6 should be understood only as requiring a substantial satisfaction of themselves. And while it may be true that there is no obvious sense in which the 'positive reasons' thesis (D6) could be satisfied in a non-substantial way, this qualification is crucial for the 'no undefeated defeaters' condition, the point being that if the hearer does not have any undefeated defeaters because she is incapable of having any at all, then she is not *praiseworthy* (justified/ warranted) for accepting an otherwise reliable testimony and, therefore, she lacks knowledge.

Now, to see how this is connected to Pritchard's account, recall that COGA$_{weak}$ reads that for *S* to know that *p*, *S*'s true belief must be the product

[89] One of the examples that motivate Lackey's view is the following:

> For instance, one of the reasons it doesn't make sense to impose a "no-lying condition" on a chair is because chairs cannot lie. To say that a chair has satisfied such a condition merely because it hasn't lied, without taking into account whether the chair has the capacity to lie, trivializes what satisfaction of such a condition. Of course, considerations of this sort apply to persons as well. (Lackey 2008, 197)

of some reliable belief-forming process that is appropriately integrated within *S*'s cognitive character, such that the cognitive success is significantly creditable to *S*'s cognitive agency. Since credit is attributed in cases of success through ability, however, this means that the employment and exercise of the belief-forming processes,[90] via which *S* came to rationally accept the speaker's testimony, must signify that *S* has exhibited some *effort* for which his/her cognitive agency is *praiseworthy*, and so, believing the truth is significantly creditable to him.[91] And this, in turn, is on a par with Lackey's demand that the acquisition of positive reasons and the failure to detect any negative ones are conditions on testimonial knowledge that must be *substantively* satisfied.

And finally, let us turn to Lackey's only condition that pertains to the speaker, namely D2: *the speaker's testimony must be reliable or otherwise truth-conducive*. First, we should concentrate on the epistemic burden distribution that the inclusion of D2 entails. As it has been previously noted, Lackey's dualism, contrary to reductionism and non-reductionism that only focus either on the hearer or the speaker, distributes the epistemic burden across both parties of the testimonial exchange. But how can COGA$_{\text{weak}}$ account for the dual origins of the epistemic justification/warrant? According to COGA$_{\text{weak}}$, knowledge can be attributed to *S* only if the cognitive success of believing the truth can be significantly credited to *S*'s cognitive agency. Crucially, however, COGA$_{\text{weak}}$ denies that the cognitive success must be *wholly* attributed to the hearer's cognitive agency thereby allowing, in cases of testimonial knowledge, for the rest of the credit to be, at least in part, attributed to the speaker's epistemic effort. To see how this would work, it should be helpful to go back to the Jenny case; it is not that Jenny's cognitive character has nothing to do with her believing the truth; it is just that the informant's cognitive character is more important. Despite this fact, however, a significant part of the credit can be attributed to Jenny's cognitive agency for

[90] Remember that, in cases of testimonial knowledge, the belief-forming processes found in the formulation of COGA$_{\text{weak}}$ stand for the inductively and memory based positive and negative reasons that one may have for rationally, or at least not irrationally, accepting, or rejecting a speaker's testimony (i.e., for appropriately integrating, or not, the speaker's reports within the rest of one's cognitive character).

[91] Notice that Lackey, not having in mind the 'of credit/creditable' debate between Greco and Pritchard, freely speaks of praiseworthiness, without, however, adopting any positive stance on the issue. Therefore, in this and the previous paragraph, 'praiseworthy' could be substituted for 'creditable' without any loss at the force of the argument. Manifesting some effort could be creditable to one, without the need to further specify whether the action performed was positive, negative, or neutral.

employing the right sort of belief-forming processes for rationally accepting the speaker's words. At the same time, however, the rest of the credit can be, at least in part, attributed to the speaker's cognitive agency for delivering a reliable report. So, we see that, in this way, COGA$_{weak}$ can accommodate the very essence of Lackey's dualism in the epistemology of testimony, namely the epistemic burden distribution across both the speaker and the hearer.

To conclude, then, COGA$_{weak}$ does appear to accommodate one of the most detailed accounts of testimonial knowledge on offer. Since testimony appears to be responsible for a preponderance of our true beliefs that amount to knowledge this should be a felicitous conclusion on its own. What is distinctive of Lackey's dualism on testimonial knowledge, however, is the fact that it explicitly points out to the dual sources of testimonial justification. That is, testimonial justification is not fully reducible to the hearer's reasons for rationally accepting the speaker's report. Instead, it also supervenes on the reliability of the speaker's report. Remarkably, COGA$_{weak}$, which is meant to capture a necessary aspect of knowledge understood in terms of creditable true belief, allows the hearer to acquire knowledge, because the cognitive success can be significantly creditable to her cognitive agency. At the same time, however, it allows for the rest of the credit to be attributed to the speaker for offering a reliable report. Accordingly, COGA$_{weak}$ has the resources to do justice to Lackey's insight regarding the epistemic burden distribution that takes place in cases of testimonial knowledge. And since knowledge is understood as creditable true belief and credit must be attributed to both parties of a testimonial exchange, testimony appears to be a type of knowledge that, in a first sense, justifies the central claim of this chapter: knowledge, in many cases, is essentially both individual and social.

4.3) Epistemic Coverage Support

In his recent book, Sandy Goldberg (2010) appears to share Lackey's insight with respect to the epistemic burden distribution that occurs in cases of testimonial knowledge. In fact, in order to accentuate the speaker's involvement in the production of a reliable testimonial belief, he goes so far as to claim that the belief-forming process that produces the hearer's justified true belief is a single belief-forming process that supervenes on both the hearer and the speaker's cognitive sub-processes: "far from being merely local features of the subject's environment, the testimony itself, along with the cognitive processes implicated in the production of that testimony, are more appropriately regarded as *part of the testimonial belief-forming process itself*. Call this the "extendedness hypothesis"" (2010, 79).[92]

In so arguing about testimonial knowledge, Goldberg's ultimate aim is similar to the aim of the present chapter, which is to make explicit the social dimension of some kinds of individualistic knowledge. Clearly, if testimonial knowledge is the product of a belief-forming process that '*epistemically* extends' from the hearer to the speaker's cognitive processes, then, in virtue reliabilistic terms, a significant part of the credit for the hearer's cognitive success should be attributed to both parties of the testimonial exchange.[93] Although Goldberg's account of testimonial knowledge is very interesting, we do not here need to dwell on its details, since the aim of this section is to concentrate on another very interesting epistemological phenomenon, which, as Goldberg himself claims, is not an instance of his "extendedness hypothesis". I am referring to the 'coverage-reliability' of one's community.

To illustrate Goldberg's point, consider that you know that there is no World-War taking place at the moment, that none of your colleagues was fired in the past few days, that there are no protests taking place at the city center right now, that Messi has not signed a contract with Real Madrid and that Madonna is not dead. One of the underlying reasons for all these instances of knowledge, Goldberg claims, is that *if any of those beliefs were true, you would have heard about it by now.* Call the italicized conditional the 'true-to-testimony conditional'. So, in some more detail, any coverage-supported belief will be a "species of inferential belief, where one of the premises involved is none other than (something like) the truth-to-testimony conditional itself" (Goldberg 2010, 174). Specifically, a subject's coverage supported belief that *p* is justified by her current belief that she has no

[92] Notice that Goldberg does not make this claim on the basis of the extended cognition hypothesis, and indeed, as formulated in the previous chapters, HEC would rule out testimony as an extended cognitive belief-forming process. Rather, Goldberg holds that testimony is a belief-forming process that *epistemically* "extends" to the cognitive capacities of the speaker because testimony is a 'quasi-belief dependent' belief-forming process whose *reliability is a function of the reliability of its input*, which, in turn, depends on the reliability of the speaker's cognitive processes that produced the relevant testimony. For more details see (Goldberg 2010, chs. 3, 4). For Goldberg's disavowal of HEC see (2010, ch. 5).

[93] Notice that Goldberg commits himself only to process reliabilism. Accordingly, his way of unraveling the social dimension of knowledge is by 'socializing reliability'.

memory of having been informed that not-*p*, *together with* her belief in the relevant instance of the truth-to-testimony conditional. There are, of course, relevant inferences that will not hold. So here is a set of five conditions that Goldberg (2010, 158-164) deems jointly sufficient for a subject's coverage-supported beliefs to count as knowledge:

i) *Source existence condition*: there must be some subgroup of members of the hearer's community—we will call this group "the source"—who are disposed to report about the relevant sort of matters.

ii) *Reliable coverage condition*: the relied-upon source must be reliable in uncovering and subsequently publicizing truths about the domain in which the subject is exhibiting coverage-reliance.

> Let *D* be a domain of interest to subject *H*, let *p* be any proposition in *D* regarding whose truth *H* might take an interest, and let α be some source on whom *H* could rely on matters pertaining to *D*. Then we can characterize the relevant notion as follows: CR is *coverage-reliable* in $D =_{\text{def}} \alpha$ (i) will (investigate and) reliably determine whether *p*, (ii) will be reliable in reporting the outcome of that investigation, and (iii) will satisfy both of the previous two conditions in a timely fashion. (Goldberg 2010, 159)

iii) *Sufficient interval condition*: there must be some sort of coordination between the time-related expectations of *H*, on the one hand, and the abilities of *α* to make any relevant discoveries, on the other:

> *α* (the relied-upon source) must be such that, at the time t at which *α* is being relied upon by *H*, it is true that, were there some relevant discovery to be made, *α* would have made the relevant discovery by t, and would have reported on the matter. (2010, 160)

iv) *Silence Condition*: in point of fact, *H* has not encountered any relevant report to date.

v) *Receptivity Condition*: *H* must be such that she *would* come across whatever relevant reports were offered by the source(s) on whom she was relying, were one to be made.

Now, letting any relevant details aside, the above conditions should be relevantly uncontroversial. What is more, notice how many of them really pertain to the subject whose knowledge status is being assessed. Remarkably, as Goldberg notes, it is only the last two conditions that refer to the hearer. What those two conditions require is that the hearer be able to remember whether she has indeed encountered any relevant reports in the past and that she is actually open to the channels of information that she is relying upon. Now, moving to the *sufficient interval* condition, it apparently refers both to the hearer and the relied upon source. The hearer, on her part, must be able to appropriately appreciate how fast her relied upon source can regularly investigate and report on the matters it is relied upon, whereas the relied upon source must conform to the time interval that is regularly considered to take place between the occurrence of a relevant fact and its publication. Finally, both the first and second conditions refer to the subject's community. There must be appropriate informational channels, which reliably both investigate and subsequently report on relevant matters.

So if the epistemic phenomenon that Goldberg has unearthed does indeed obtain—and it appears that it does—one obvious conclusion is in order. Since of the five jointly sufficient conditions that must be satisfied in order for a subject's belief to be coverage-supported such that it can amount to knowledge, two of them pertain to the hearer's cognitive abilities, in such cases, the hearer's cognitive success will be significantly creditable to the hearer's cognitive agency. According to $\text{COGA}_{\text{weak}}$ then, a subject can gain knowledge on the basis of coverage support. What is of more interest, however, is that one of the five conditions is shared by both the hearer and her relied upon source, whereas the remaining two conditions really pertain to the epistemic community that the hearer is embedded in. Accordingly, the remaining epistemic credit should be attributed to the hearer's epistemic community. And since, as it has been previously noted, knowledge is here understood in terms of creditable true believing, coverage-supported true beliefs is another instance of knowledge which is at the same time both social and individual.

4.4) The Individual, Epistemic Artifacts, and the Society

Testimonial beliefs and coverage-supported beliefs demonstrate the social nature of individualistic knowledge as in both cases the subject's cognitive success is not only creditable to her cognitive agency but also to other individual epistemic agents who have either contributed to the reliability of the testimonial reports, or the coverage-based beliefs that the subject rationally accepts as true. These are straightforward manifestations of the claim that individual knowledge can at the same time be social in the sense that the process of knowledge acquisition pertains to considerations that clearly go beyond the subject of knowledge. In this section, however, the aim is to concentrate on another category of cases whereby, even though one's epistemic society affects one's knowledge-acquisition only indirectly, individual knowledge may again turn out to be a social phenomenon as well.

The cases I have in mind are none other than the cases in which the agent comes to know something on the basis of the operation of an epistemic artifact. As it has been previously noted, in such cases, the cognitive success will be *significantly* creditable to one's cognitive agency, for it is one's cognitive agency that is first and foremost responsible for the integration and sustainment of the extended belief-forming process so as to eventually accomplish the aim of forming a true belief on the relevant matter.

In contrast, there appear to be cases of perceptual knowledge whereby one could claim that the credit can be *wholly* attributed to the agent's cognitive agency, as he has solely employed his organismic cognitive faculties in order to form his true beliefs. This, however, may be a simplification. For if we consider the theory-laden nature of empirical observations and the other underlying processes that influence our understanding of the surrounding world, then, again, the cognitive agent seems to employ more processes than merely the ones that make up his organismic cognitive mechanism.[94] I will leave the question on the nature of perceptual knowledge open for future research.

Now, to return to cases whereby an epistemic artifact is explicitly involved in the process of knowledge acquisition, one may fairly wonder whereto should the rest of the credit be attributed? This is a fair worry, for as

[94] The main idea here is that perceptual knowledge, i.e., the output of perceptual abilities, is observational. We also know, however, that observation is theory-laden, either by explicitly articulated scientific theories, or by implicit "natural interpretations" (Feyerabend 1975, ch. 6). Now, if we accept the claim that scientific theories are themselves languages, which according to HEC are the ultimate (software) artifacts, then we are also led to the conclusion that even individual observations—the output of perceptual abilities—are not strictly individual in nature. Scientific theories are socially constructed.

it has been argued in the previous chapters, in such cases, the prevailing factor in the causal explanation of the agent's cognitive success is the *integrated* extended belief-forming process that consists of both one's cognitive agency and the epistemic artifact, operating in tandem. So, since a significant part of the credit has been attributed to the agent's cognitive agency, should we attribute the rest of the credit to the relevant external reliable belief-forming process? That is, should we attribute credit to telescopes, microscopes, calculators, languages, scientific theories and so on? It seems that the answer to these questions should be negative.

To see why, consider that even though Greco (2004) holds that credit attributions are very much akin to causal explanations, attributions of responsibility, praise, or merely neutral action (i.e., attributions of positive, negative, or merely neutral credit, respectively) have been traditionally associated with *intentional* agents. Since, however, it would be highly implausible to claim that artifacts have intentions, it follows that no credit (be it positive, negative, or neutral) could be attributed to them.[95] Notice, nevertheless, that artifacts can be defined as objects that have been *intentionally* made or produced for a certain purpose. Accordingly, what the above points seem to suggest is that even though no credit can be attributed to artifacts, whenever they figure most importantly in causal explanations, the corresponding credit could be attributed to the individuals that produced the relevant artifacts. Credit, therefore, cannot be attributed to telescopes, microscopes, calculators, languages and scientific theories; rather, in cases where knowledge is the product of an extended belief-forming process, the rest of the credit should be attributed to the individuals that brought the relevant belief-forming processes about. Mind, however, that, frequently, we will not be able to attribute the rest of the credit to only one single individual, because, in most cases, in order to come up with such reliable belief-forming processes the individual must employ similar belief-forming processes, or rely on knowledge that has been delivered by other individuals on the basis of further reliable belief-forming processes and so on. Now, before drawing

[95] In so avoiding to determine whether the credit will be negative, positive, or neutral, the issue of where it could be attributed to is orthogonal to Pritchard and Greco's debate with regards to the question of whether knowledge is always 'of credit', and thereby an achievement.

our conclusions, let us consider two examples that may render the present view more intuitive.

Consider the FIA Formula One Championship. The F1 season consists of a series of races, the results of which are combined to determine *two* Annual World Championships; one for the drivers and one for the constructors. In this case, the analogue to be drawn is that the drivers play the role of the cognitive agents and the cars that of the epistemic artifact.[96] According to FIA's rules, however, the credit for winning cannot only be attributed to the drivers; hence the two championships. Moreover, pay attention to the fact that the second championship is not attributed to the cars but to the constructors that built the cars. In other words, the credit for winning is not only attributed to the cognitive agent that drives the car, but, also, to the *team* that brought about the racing artifact (i.e., the car).

Next, consider a meteorologist who advises the sailors not to travel tomorrow because he knows there will be a storm. For coming to know the target proposition (that there will be a storm tomorrow) the meteorologist collects weather observations of atmospheric pressure, temperature, wind speed and direction, humidity, etc., which he inserts as inputs to a supercomputer. The supercomputer performs a simulation of the atmosphere, generating outputs, usually in the form of graphs, which then the meteorologist studies, feeds back new inputs in the supercomputer, and eventually forms the belief that there will be a storm. In other words, the meteorologist reciprocally interacts with the computer such that he appropriately integrates within his cognitive character the belief-forming process of the computer simulation so as to come to know the truth of the target proposition.

Obviously, the meteorologist could not have come to know the target proposition soon enough—though one could argue that he could not have come to know it not even within a lifetime—without employing the reliable belief-forming process of the computer simulation. That is, for coming to know the target proposition, the meteorologist has to rely on an extended belief-forming process that was brought about on the basis of knowledge—produced by further reliable belief-forming processes—of long generations of mathematicians, computer scientists, chemical and electrical engineers,

[96] Driving the car then, plays the role of the overall extended belief-forming process.

physicians, and, in general, a vast series of experts whose length could go on for a while.[97] Had the meteorologist not been a part of this epistemic social structure, and therefore lacked the necessary reliable belief-forming process, he would be incapable of gaining knowledge of the target proposition.

Overall, then, the cognitive success of coming to know there will be a storm is to a significant degree creditable to the particular meteorologist—it is he who came to know the target proposition by employing the necessary belief-forming process—but the rest of the credit must be attributed to the individuals, and in general to the social structure, that brought about the necessary belief-forming process.

Furthermore, consider how a similar description of the process of gaining propositional knowledge could apply within the fields of astronomy, mechanics, physics, economics, biology, chemistry, neuroscience, mathematics, and so on. The moral seems to be that in order for one to know something on most of the matters, one needs the whole historical and cultural background that can generate and complement one's expert knowledge. One needs the background knowledge and the reliable belief-forming processes produced by long generations of mathematicians, engineers, experimentalists, scientists, traders, explorers, philosophers and many other experts so as to come to know the truth of some proposition *p*. (Remember that, in section 3.4, it was argued that even scientific theories may qualify as extensions of one's cognitive character). In other words, one could argue that the individual agent may have an advanced epistemic standing only within a given social structure.[98]

Now, to pick up our epistemological discussion from where we left it, in the cases above, the knower *is* the individual who comes to know some proposition *p* by employing some reliable belief-forming process, which is appropriately integrated within her cognitive character such that the

[97] Notice, further, that the observations which the meteorologist uses as inputs to the computer simulation are also the product of instruments, or reliable belief-forming processes produced by a series of individuals who have relied on the knowledge grounded on further reliable belief-forming processes and so on.

[98] By 'advanced epistemic standing' I mean any kind of knowledge beyond perceptual knowledge. That is, knowledge that in order to be obtained, the individual agent has to employ extended belief-forming processes, or evidence delivered by extended belief-forming processes, or testimonial reports. Notice, however, that considering the theory-laden nature of the empirical observations and the rest of the underlying processes of our understanding of the world, one could strengthen the point by saying that the individual may be an epistemic agent, *in general*, only within a given social structure.

cognitive success is to a significant degree creditable to her cognitive agency. Crucially, however, the epistemic agent has come to know some proposition *p* by appropriately integrating within her cognitive character an extended reliable belief-forming process. In these cases, it has been argued, the rest of the credit for the cognitive success of *S*'s believing the truth with regards to the target proposition will have to be attributed to the individuals, and, in general, to the social structure that brought about the relevant reliable belief-forming process.[99] Therefore, since, according to the broader framework of virtue reliabilism, knowledge is creditable true belief that is the product of cognitive abilities, knowledge that is the product of the operation of epistemic artifacts turns out to be, essentially, both social and individual.

In discussing the embodied and situated nature of cognition, Gallagher writes: "I do not disagree with Dennett concerning the role played by non-conscious elements, except that I think we are even larger than he thinks—we are not just what happens in our brains. The 'loop' extends through and is limited by our bodily capabilities, into the surrounding environment, which is social as well as physical, and feeds back through our conscious experience into the decisions we make" (Gallagher 2005, 242). We, thus, see that cognition is both embodied and physically and socially situated. But, note that if knowledge is the product of cognitive abilities, then so knowledge, too, turns out to be both embodied and physically and socially situated.

In other words, according to virtue reliabilism and, in particular, to $\text{COGA}_{\text{weak}}$, we gain a view on knowledge whereby the individual agent can be an advanced epistemic agent only within a given social structure necessary for supplying him with reliable-belief forming processes that he later integrates within his cognitive character so as to come to know the truth of

[99] Conversely, if the external belief-forming process is not a reliable belief-forming disposition that is appropriately integrated within the agent's cognitive character, then the agent won't be credited with knowledge, and, thereby, neither the rest of the social structure that brought about the relevant belief-forming process will. For example, if when using a telescope for the first time I see something that looks like a star, I cannot be credited with knowledge of the proposition that 'there is a star', because what I see could have so easily been something completely different due to observational noise produced by, say, the atmospheric turbulence. In this case, the extended belief-forming process of telescopic observation is not appropriately integrated within my cognitive character, as it is not a belief-forming disposition for me (I use the telescope for the first time). Therefore, I cannot be credited with knowledge of the target proposition and, thereby, neither the rest of the social structure can. That is, the social structure can be credited with knowledge only through the epistemic agent's cognitive agency that has come to know the truth of some proposition by appropriately integrating within his cognitive character the external belief-forming processes produced by the social structure.

some proposition *p*. At the same time, however, the social structure can gain knowledge and produce the relevant reliable belief-forming processes only through the necessary epistemic agency of the individual cognitive agent. This interesting epistemic interaction renders some more plausibility to the intuitive idea that the society and the individual are the two aspects of the same entity and, thereby, the one cannot be considered in isolation from the other.

4.5) Conclusion: Epistemic Dependence and Epistemic Individualism: the Dual Nature of Knowledge

It has been argued so far that individual knowledge, understood along the virtue reliabilistic lines as creditable true believing, so often appears to be also social in nature. In particular, the focus has been on cases of testimonial knowledge, knowledge of coverage-supported beliefs, and knowledge acquired on the basis of the operation of epistemic artifacts. These cases admittedly represent a very important fraction of a subject's channels of knowledge acquisition. In those cases, it was argued, the subject's cognitive success is *significantly* creditable to her cognitive agency and thus, according to COGA$_{weak}$, the individual is knowledgeable. Interestingly, however, the rest of the credit, in all these cases, is creditable to the individuals that belong to, and form the epistemic society, or culture, in which the subject of knowledge is embedded. This is so, either because those individuals offer reliable reports, form epistemic channels on which the subject can rely on for her coverage-supported beliefs, or have produced some reliable belief-forming process that the subject can integrate within her cognitive character so as to reliably form true beliefs. Therefore, to repeat the claim, if knowledge is to be understood in terms of creditable true believing and significant credit must be attributed both to the individual subject and the epistemic society (or aspects of it) of which the subject is a proper part, then individual knowledge turns out to be in many cases social as well.

But, does this partly social nature of knowledge suggest, as some epistemologists have argued, that we should stop considering the individual as the proper object of our epistemological inquiries? To answer this question, let us briefly go through what Hardwig says concerning the following case:

A knows that *m*
B knows that *n*
C knows (1) that *A* knows that *m*, and (2) that if *m*, then *o*
D knows (1) that *B* knows that *n*, (2) that *C* knows that *o*, and (3) that if *n* and *o*, then *p*.

Having this case in mind, Hardwig writes: "Suppose that this is the only way to know that *p* and, moreover, that no one who "knows" that *p* knows that *m*, *n* and *o* except by knowing that others know them" (1985, 348). On the face of these considerations, Hardwig concludes that unless we do not want to retain that scientific research and scholarship result in knowledge, because of their cooperative methodology, we must support that *p* is known in this way. However, if *p* is so known, Hardwig further claims, then it is not known by anyone person but by the community that consists of *A*, *B*, *C* and *D*, because "this community is not reducible to a class of individuals, for no one individual and no one individually knows that *p*" (1985, 349).

This is indeed a threatening conclusion for the individualistic nature of knowledge and a rather counterintuitive one. For this reason, let us try to investigate what may have gone wrong with Hardwig's argument, leading him to this dubious statement. First, we must straighten out the fact that Hardwig considers the classical account of knowledge as 'justified true belief', whereby justification is internally conceived. That is, Hardwig considers that the individual *S* must be able to justify his beliefs only by reflection alone; one of the weakest formulations of internalist justification holds that *S*'s reasons for his true beliefs must be *accessible* to him by reflection alone.[100] Accordingly, even this weak formulation of internalism leads to the conclusion that epistemic agents must be intellectually autonomous. Therefore, considering intellectual autonomy as a prerequisite for knowledge and having in mind the above cooperative process for gaining knowledge, Hardwig validly draws the counterintuitive conclusion that, in most cases, we should claim that the seat of knowledge is the community and not the individual agent. Is there, however, an alternative way to go so as to avoid this devastating conclusion for individual knowledge?

[100] I am referring to accessibilism, which is also supposed to be the standard internalist view. Roughly stated: Whenever one knows that *p*, then one can become aware by reflection of one's knowledge basis for *p*. For more details see (BonJour 1980; Chisholm 1977; Steup 1999).

It seems that, instead of challenging the individual nature of knowledge, one could bring into question the classical definition of knowledge as 'internally justified true belief'.[101] It is a happy incident, then, that virtue reliabilism and COGA$_{\text{weak}}$ are externalist approaches to knowledge that have been proposed as alternatives to the classical internalist account, leading to the more intuitive conclusion that the individual *is* a proper epistemic agent, even though not autonomously so. To see how this might be so consider that, according to externalism, in order for one's true beliefs to count as knowledge, they need not be backed up by reasons to which one could in principle have introspective access. And so, denying the demand for introspective access to one's justification for one's beliefs also makes the demand for intellectually autonomy appear undermotivated. COGA$_{\text{weak}}$, however, takes these considerations a step further. More in particular, by allowing knowledge to be acquired merely on the basis of reliable belief-forming processes such that the cognitive success can only be *significantly* creditable to one's cognitive agency it actually anticipates—if not ascertains—the denial of intellectual autonomy. Either the reliability of the input for, or the reliability of the very belief-forming processes themselves may heavily depend on one's epistemic society/culture and this is a fact that can be explicitly accommodated; COGA$_{\text{weak}}$ allows (the rest of the) credit to be attributed to those exogenous (or rather extra-organismic) epistemic factors as well. Nevertheless, at the same time, it stresses the importance of the individual epistemic agent by demanding that the cognitive success be significantly creditable to his/her cognitive agency.

To be clear, however, mind that the above is not to claim that all externalist epistemology points away intellectual autonomy and *robust individualism*—as we may call the view that knowledge should be fully down to the individual. Having the reasons of one's justification out of one's awareness is certainly not the same as partly having those reasons out of one's bodily boundaries. There certainly can be many externalist conditions on knowledge, which are individualistic in nature. Take for example Greco's virtue reliabilism, which demands that the cognitive success be *primarily* creditable to *S*'s cognitive character. On a first look, this proposal appears to

[101] That is, one could bring into question one of Hardwig's premises.

be a form of *weak individualism*.[102] We therefore see that the internalism/externalism distinction is by no means the same as the individualism/anti-individualism distinction. Moreover, even though internalist conditions on knowledge may always be tied to robust individualism, externalist conditions appear to come in degrees, with the potential to lie anywhere on the continuum that the individualism/anti-individualism distinction defines. Nevertheless, COGA$_{\text{weak}}$—with its lenient demands on the creditability of the cognitive success to one's cognitive agency—seems to be able to capture the full spectrum of the individualism/anti-individualism continuum. So far, we have been liberated from the demand for intellectual autonomy and the concomitant robust individualism, by pointing out that knowledge may be creditable both to the individual and the society of which he/she is a part. This might be called a version of *weak epistemic anti-individualism*. More on the most liberal—and diametrically opposite to robust individualism—version of (epistemic) externalism will be discussed in the chapters to follow.

For now, it suffices to conclude that knowledge can be both social and individual in nature exactly because, in most cases, the individual comes to know the truth of some proposition *p* by relying on the support provided by the social structure—through either the employment of socially derived belief-forming processes, the society's *epistemic structure*, or by exploiting information that has been delivered by other individuals on the basis of further reliable belief-forming processes and so on.

[102] If, instead, it demanded that one's cognitive success be *solely* creditable to one's cognitive agency, it would be a case of robust individualism.

CHAPTER 5

The Hypothesis of Distributed Cognition

5.1) Introduction

In the previous chapter we focused on testimonial knowledge, coverage-supported beliefs, and knowledge acquired on the basis of the operation of epistemic artifacts. In these cases, even though the agent's justification lies partly outside his cognitive agency (i.e., his organismic cognitive apparatus), the agent manifests sufficient cognitive effort such that the cognitive success is significantly creditable to him. Thus, on the basis of COGA$_{weak}$, we concluded, even though the agent is not intellectually autonomous, knowledge can still be properly ascribed to him.

This weak version of anti-individualism also points towards the dual nature of knowledge. Knowledge—understood as creditable true belief—is individualistic in such cases, because the agent manifests cognitive effort such that the cognitive success is significantly creditable to him. In the case of testimony and coverage-supported beliefs, the cognitive success is significantly creditable to the agent because on the basis of his reliable cognitive abilities he either rationally accepts the speaker's reliable report, or checks that he has not come across a contrary piece of information. In the case of knowledge attained on the basis of artifacts, cognitive success is again significantly creditable to him, because it is his cognitive agency that is first and foremost responsible for recruiting, maintaining, and overall integrating within his cognitive character the artifact that partly constitutes the extended belief-forming process on the basis of which he acquires a true belief. But, in all of the above cases knowledge is social as well. This is because in testimonial and coverage-supported beliefs, part of the credit should be attributed either to the testifiers, or to the individuals that bring about the socio-epistemic structure that allows the epistemic agent to enjoy coverage-supported beliefs. Similarly, in the case of epistemic artifacts, part of the credit should be attributed to the individuals who brought those artifacts about. Accordingly, the above ways of knowing demonstrate that knowledge may be partly social while being, at the same time, individual as well. Therefore, in such cases of epistemic dependence, we do not need to embrace Hardwig's unpalatable conclusion, namely that the known proposition is not known by anyone individually, but only by the society as a whole. Going against robust individualism does not necessarily entail full-blown anti-individualism. The middle grounds of weak anti-individualism are open for us.

Yet, I here wish to examine whether Hardwig's conclusion may still be true. That is, could there be cases of knowledge where it is not known by anyone individually? In other words, could we make a case for robust anti-individualism in epistemology?

According to COGA$_{\text{weak}}$ and its concomitant understanding of knowledge as creditable true belief, robust anti-individualism can only get off the ground if there are cases of knowledge where the cognitive success won't be significantly creditable to any individual alone. That is, we need a case where justification will be entirely distributed among several individuals and/or epistemic artifacts. In particular—cashed out in virtue reliabilistic terms—we are after a case of distributed cognitive ability (i.e., a distributed belief-forming process) on the basis of which knowledge may be acquired, such that the cognitive success won't be creditable to anyone in particular but, instead, to a group of individuals as a whole.

In order to make such a case, therefore, we must focus on group agents, and particularly *epistemic* group agents. If such entities are possible then knowledge, which is only creditable, and thus possessed only at the group level should be possible as well. Again, being able to locate such a case would amount to making a case for epistemological robust anti-individualism.

Accordingly, this and the following chapter will focus on group agents in general and epistemic group agents in particular. More specifically, the present chapter is dedicated to the metaphysical support of group agents on the basis of emergentist considerations. Interestingly, as we shall see, these considerations will invite us to re-invoke the arguments from dynamical systems theory (DST), which were offered in support of the postulation of coupled systems in chapter 2. Once metaphysical support is provided—revealing some interesting, yet unsurprising, generic properties of group agents—chapter 6 will focus on *epistemic* group agents. In particular, I will provide two examples of alleged epistemic group agents that seem to satisfy the generic properties to be revealed in this chapter. Then, I will move on to examine whether COGA$_{\text{weak}}$ can account for such cases and whether it can further reveal some of the more specific properties of epistemic group agents in particular.

5.2) Group Agents

5.2.1) Introduction

The existence of group/collective agents has lately started receiving growing attention within philosophy of mind and cognitive science, and several approaches have been offered for arguing in their support. For example, List & Petite (2010) argue on the basis of collective judgments, Tollefsen (2002) has been interested in collective intentionality, while Theiner et al. (2010), Wilson (2005), and Hutchins (1995) focus on distributed cognitive/computational systems. Since, however, the focus here is on epistemic group agents, while also presupposing the validity of the ability intuition on knowledge—i.e., that knowledge must be the product of cognitive ability—we will only concern ourselves with considerations pertaining to distributed cognition. Remember, in order to make the case for robust anti-individualism we need to see whether there could be any distributed belief-forming processes, which are only possessed by groups of individuals and not by any individual alone. In such cases, knowledge will only be attributable to the group level.

For doing so I will rehearse the points and arguments from (Sawyer 2001), where the author claims that the best way to support the existence of social entities—or at least social properties—is the same way in which non-reductive materialism is defended within philosophy of mind. The discussion will make apparent the importance of dense non-linear interactions amongst individuals in cases of non-reducible (i.e., emergent) social entities. Thus, I will argue, there can also be a second way for arguing for group agents; namely, by re-invoking the considerations from DST presented in chapter two. Having made explicit the connection between these two argumentative lines, I will then conclude by going through further theoretical support available in the literature, which is in fact very much in line with the aforementioned arguments.

5.2.2) Supervenience, Multiple Realizability, and Wild Disjunction

Sawyer (2001) begins his article by explaining that the relationship between the individual and the collective is one of the most troubling issues in sociology, having perplexed the minds of such figures as Durkheim, Marx, Simmel, and Weber. In more liberal terms the question has been this: what is the connection between the micro and the macro?

Explanations of this micro-macro link have very often appealed to the notion of *emergence*: collective phenomena are collaboratively created by individuals, but may not be reducible to them. Furthermore, the debate on the micro-macro link has given rise to two main strands within sociology: 1) methodological individualism and 2) methodological collectivism. Methodological individualism is the claim that there exist emergent social properties, yet they can be reduced to explanations in terms of individuals and their relationships. Methodological collectivism, on the other hand, is the claim that collectives possess emergent properties that cannot be reduced to individual properties (or explanations in terms of individuals and their relationships).

As Sawyer observes, however, both of these two very distinct sociological accounts of the micro-macro link appeal to the concept of emergence. Yet, as we can see, they produce two opposite conclusions. Sociologists, however, Sawyer seems to suggest, are not philosophers; this awkward situation may result from a misunderstanding of what emergence is supposed to be. Thus, he turns to the best available philosophical account of the notion—the one arising from the arguments in favor of non-reductive materialism; i.e., the third path between dualism and identity theory within philosophy of mind.

To start with a bit of history, Sawyer notes that modern notions of emergentism were invoked in the early 20th century in order to reject vitalism and dualism, while accepting the materialist ontology that only matter exists:

> Higher level entities and properties were grounded in and determined by the more basic properties of physical matter; this was referred to as

> Supervenience. However, the 1920s emergentists argued that when basic physical processes achieve a level of complexity of an appropriate kind, genuinely novel characteristics emerge; these emergent higher-level properties could not, even in theory, be predicted from a full and complete knowledge of the lower-level parts and their relations. Further, they could not be reduced to properties and their relations, even though those properties are supervenient on and thus determined by the system of parts ((Sawyer 2001, 554), citing (Kim 1993, 134) and (Teller 1992, 140-42)).

Sawyer explains that these ideas revived within philosophy of mind in the 1960s after the rejection of behaviorism. At that time, emergence was offered as a third alternative to the debate between identity theory and dualism; the former held that the mind is nothing more but the biological brain, while the latter held that mind and brain are fundamentally distinct: the one is material while the other is immaterial. Emergence, however, gave rise to non-reductive materialism, according to which even though only physical matter exists, mental properties are not reducible to physical ones (Davidson 1970; Fodor 1974) and may indeed have causal powers over the physical brain (Anderson et al. 2000; Heil & Mele 1993).

The argument for non-reductive materialism begins with the Supervenience thesis: all there is is physical matter and all higher-level properties supervene on the system of lower-level components. Notice, also, that supervenience refers to the relation between two levels of analysis; if two events are identical with respect to their descriptions at the lower-level, then they cannot differ at the higher-level. Moreover, if an entity changes at a higher level it will also change at the lower level.[103] Obviously, however, supervenience is not enough for emergence as it is compatible with the claim that all higher-level properties are identical with lower-level ones. As Fodor (1974) argued, what is further required are the phenomena of multiple realizability and wild disjunction.

According to Fodor's argument for non-reductive materialism, a law is a statement within which the basic terms are natural kinds of that science. In order to reduce a law to the science of the lower-level, what is required is a bridge law that translates that law. But in order to come up with a bridge law,

[103] Of the last two sentences, I am only reluctant with respect to the first one. The reason is this. Even if we admit that the lower-level description may refer both to the properties of the lower-level components *and their relations*, given the discussion to follow, there could be certain non-linear relations amongst components, which may only be describable at the higher level. Accordingly, there could be identical descriptions at the lower level, which may nevertheless give rise to events that are not identical at the higher level.

what is further required is that each of the natural kind terms of the higher-level science be translated into natural kind terms of the lower-level science. The main point of Fodor's argument is that there are no a priori reasons to believe that this translation will always be possible for any given pair of scientific disciplines. A simple translation from psychology to neurobiology, or, in our case, from sociology to psychology may not be possible and this can only be empirically determined.

Now, the reason why this translation may be impossible has to do with the phenomenon of multiple realizability; even though each mental state is supervenient on some physical state, each given instance of that mental state might be realized by a different physical state. Again, however, multiple realizability alone does not entail irreducibility. If there are only a few realizing states, or if those states display some common features, the reduction may be performed unproblematically.

Nevertheless, reduction would indeed be problematic if the "neurobiologically equivalent of a psychological term were an otherwise unrelated combination of many neurobiological concepts and terms. Fodor termed such a realization wildly disjunctive" (Sawyer 2001, 557). According to Fodor, a true scientific law cannot have wildly disjunctive components, and thus, wild disjunction implies that there could be lawful relations among events described in psychological language, that would not be lawful relations in the language of physics. Take for instance the psychological law that pain leads to screaming. Pain, however, may be realized by several states (C-fibers, D-fibers, E-fibers, F-fibers, …) that have nothing in common. The same may also hold for the mechanisms that produce screaming (although this is not necessary for the argument to go through) (see figure 1). If that is the case, then no law can be found at the neurobiological level, such that it can capture the scope of the psychological law. (That is, the neurobiological properties of the several realization bases that are grouped under the same psychological kind (e.g., pain) may have nothing in common. Thus, neurobiology alone cannot group them as one single kind. These realization bases can only be grouped as such from the psychological perspective).

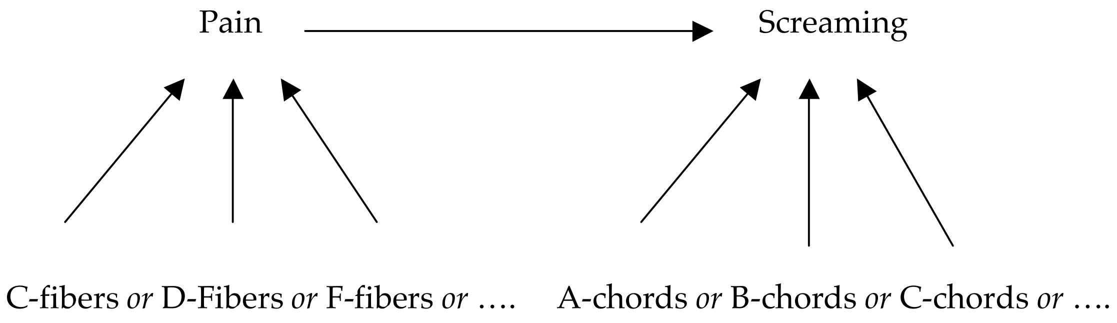
Pain
Screaming
C-fibers *or* D-Fibers *or* F-fibers *or* ….
A-chords *or* B-chords *or* C-chords *or* ….

Figure 1

Now, as Sawyer points out, even if one does not accept Fodor's understanding of a scientific law, "it is clearly of limited scientific usefulness to have laws with wildly disjunctive terms, because they provide only limited understanding of the phenomena; they are of limited predictive usefulness, because they apply only to a specific token instance, whereas the higher-level law is likely to be more generally applicable" (Sawyer 557). Thus, Sawyer goes on, "when supervenience is supplemented with the argument from wild disjunction—the observation that a single higher-level property can be realized by many different low-level supervenience bases and that these different supervenience bases may have no lawful relation with one another—we have an account of emergence that shows why certain social properties and social laws may be irreducible" (*ibid*.).

Furthermore, according to Sawyer, many social properties seem to work this way. Take for instance the property 'being a church'. Each individual member of the church may instantiate the property 'believing in X_n' such that the sum total of such individual beliefs are constitutive of the social property 'being a church'. Yet, a wide range of individual beliefs may realize the property 'being a church'. The same, Sawyer claims, is true of properties such as 'being a family', 'being an argument', 'being a collective movement' and so on.

Now, what is even more interesting is that the above analysis may also provide an explanation of downward causation; the claim that higher-level properties may have causal powers over lower-level properties. To see how this might be so, consider that "there can be a lawful, causal relation between a mental property and a physical property, even though there is no lawful, causal relation between the realizing physical properties of that mental property and the caused physical property" (Sawyer 559). So, for instance, in the previous example, the higher-level mental property of 'being in pain' may cause any of the realizing physical states of the physical property 'screaming' while there being no lawful, causal relation between the properties of the several realization bases of 'being in pain' and those of 'screaming' (see figure 2).

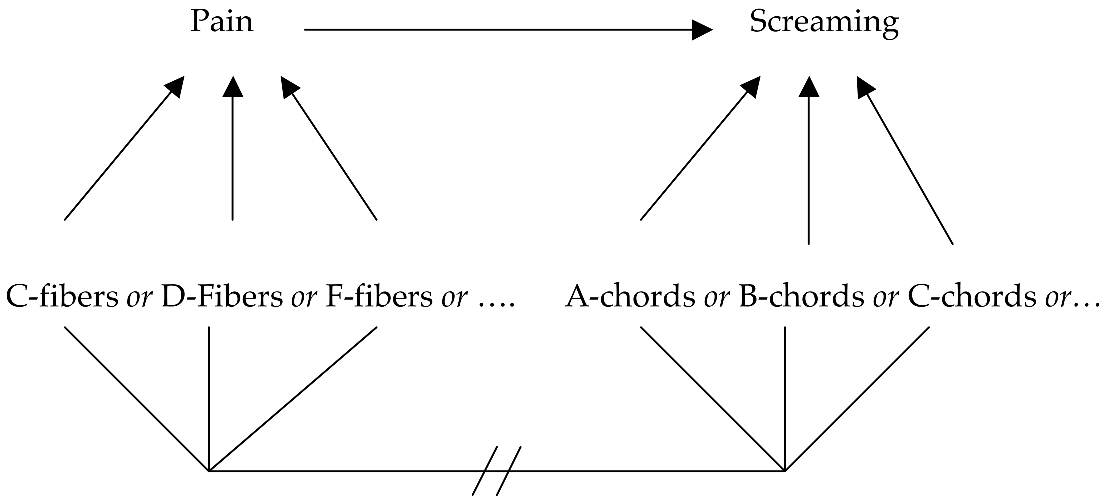
Pain
Screaming
C-fibers *or* D-Fibers *or* F-fibers *or* ….
A-chords *or* B-chords *or* C-chords *or*…

Figure 2

Accordingly, mental causation *is* a lawful relation between a mental property and a physical property, even though the causal force of the mental property inheres in its physical level supervenience base. Similarly, we might claim, social causation *is* a lawful relation between a social property and an individual one, even though the causal force of the social property inheres in its individual level supervenience base. In this sense, mental/social properties do constrain matter/individuals even though, at the same time, they are supervenient on the actions and interactions of those very materials or individuals.[104]

[104] Notice that Jaegwon Kim (1989) denies that there could be downwards causation on the basis of his 'explanatory causal exclusion' principle. That is, "no event can be given more than one *complete* and *independent* explanation" (Kim 1989, 79). So, in the example just provided, if, for some token instance of the occurrence of the two events, there is a sufficient causal relation between the underlying physical realization bases of the two events (between, for instance, C-fibers and A-chords), then it will be redundant to claim that the mental property of pain (that supervenes on its physical realization basis) also caused the A-chords vibration (i.e., screaming). In particular, the mental property of being in pain will be epiphenomenal. Therefore, Kim further argues, since mental properties have no causal force they are not real.

We thus see that wild disjunction provides a strong argument in favor of non-reductive emergentism and downward causation. In fact, the possibility of downward causation should further be an indication of the reality of the higher levels of analysis. If we can identify that a phenomenon has causal power, then we must treat it as real. Despite all that, however, Sawyer draws a somewhat conservative conclusion for the social level. Drawing the analogy from non-reductive materialism, which holds that even though only matter exists, there also are emergent mental properties, he writes: "the analogous position in sociological theory would be to hold that only individuals exist and that social entities do not have distinct existence, yet there may be irreducible social properties and social laws" (2001, 559). In other words, this amounts to rejecting sociological realism, while supporting methodological collectivism (Sawyer 2001, 552). This is a thesis, however, that strikes me as quite peculiar. How could there be properties at some level and yet possessed by no entities at that level?

As a brief explanation, what seems to be going on here is that Sawyer, having drawn the analogy from non-reductive materialism, which holds that nothing else other than matter exists, wants to say that nothing else other than individuals exist. Yet, what seems to be wrong here is that the non-reductive materialist would not want to deny the existence of minds (such minds, however, would still be material). If he did, he may equally have to deny the existence of brains, while holding that neural/brain properties exist—and notice here that the supervenience of brain/neural properties on chemical/physical properties is far more obvious than that of mental on neural/brain properties. In so far as brain/neural properties exist, brains exist. Similarly, in so far as mental properties exist, minds exist. Granted, those minds supervene only on matter, and maybe only on brains—though strictly speaking, given the extended cognition hypothesis, they need not be. Accordingly, it seems that the correct analogy to be drawn is that in so far as social properties exist, social entities exist as well. Again granted, those social entities supervene only on matter, and maybe only on individuals—although again, given the extended and distributed cognition hypotheses, they need

To see how the present analysis may solve this problem see fn. 105. Kallestrup (2006), however, has argued that the non-reductive materialist does not necessarily have to account for downwards causation.

not be, and as we shall see they arguably don't. I will return to the issue of the existence of higher-level entities at the end of the following section.

5.2.3) Characteristics of Wild Disjunction-Dynamical Systems Theory

To be clear about the main argument so far, let us repeat the claim by quoting Sawyer (2001, 564-565) in length:

> A plausible account of an instance of the micro-to-macro emergence of a social property may be provided, but that account might not be the one that actually led to the emergence of that property in that token instance; due to multiple realizability the instance of the social property being modeled might have emerged from a different supervenience base. But suppose that all agree that the micro-to-macro emergence of a social property has been successfully modeled for one token instance of that social property. This still leaves us with the second more foundational problem: that account may not be applicable for any other token instances of the same property, due to multiple realizability and it may not provide any explanatory power beyond that token instance, due to wild disjunction. […] [Moreover] if social properties are implemented in wildly disjunctive sets of individual properties, then social terms and laws may not be lawfully reducible to individual terms and laws. If a social property has a wildly disjunctive individual base, then the social property can participate in causal laws even though there is no equivalent lawful description in the language used to describe individuals.

We thus see that the argument from wild disjunction can provide a strong case for emergence. Notice, however, that it does not demonstrate that all social properties will be emergent. Whether a higher-level property, or law is not fully reducible to the lower-level is a matter that can only be empirically determined. It should be very interesting, then, to see when wild disjunction is likely to occur. For doing so, Sawyer suggests that we should focus on the *temporal mechanisms and processes* of emergence that give rise to social properties. He also notes that in order to argue in favor of emergence, theorists have mainly stressed the importance of the *interactions* of the individuals and their properties, as opposed to the properties of the individuals themselves. "Interaction is central; higher-level properties emerge from the interactions of individuals in a complex system" (Sawyer 2001, 574). Yet, since reductionism is defined as the ability to predict higher-level properties from knowledge of the lower level parts *and their relations/interactions*, it appears that an unqualified appeal to interactions is problematic. Nevertheless as we shall see, the answer may lie at the complexity of interactions between individuals, or, more generally, components of a system.

As Sawyer notes, in the late 1980s and 1990s complex system theorists began to identify the characteristics of systems that could be realized in a wildly disjunctive way. The first characteristic is *non-aggregativity*, and has been offered by Wimsatt (1986) who identified emergence with the failure of aggregativity. Aggregative properties, according to Wimsatt, meet four criteria: 1) The system property is not a product of the way the system is organized; the parts are intersubstitutable without affecting the system property. 2) An aggregative property should remain similar under addition or removal of a part from the system. 3) The systemic property should remain invariant under operation of decomposition and reaggregation of the parts. Finally, 4) there should be no cooperative or inhibitory interactions among the parts of the system for this property. Aggregativity thus defined means that a system's failure to satisfy any of the above conditions with respect to some of its properties will signify that the relevant properties are emergent. Yet, as Wimsatt himself claims, this account is too broad, and in this sense most social properties will be emergent. Moreover, Wimsatt notes that this understanding of emergence is compatible with reductionism and so cannot speak to the heart of the debate under consideration.

A second characteristic of emergent properties that Sawyer provides is *non-decomposability*:

> Decomposable systems are modular, with each component acting primarily according to its own intrinsic properties. Each component is influenced by the others only at its inputs; its function (processing of these inputs) is not itself influenced by other components (Simon 1969). In such a system, the behavior of any part is *intrinsically determined*: it is possible to determine the component's properties in isolation from the other components, despite the fact that they interact. The organization of the entire system is critical for the function of the system as a whole, but the organization does not provide constraints on the internal functioning of components. […] In contrast, in non decomposable systems, the overall system organization is a significant influence on the function of any component; thus component function is no longer intrinsically determined. Dependence of components on each other is often mutual and even make it difficult to draw firm boundaries between components (Sawyer 2001, 577).

And this, in turn, leads quite naturally to the third characteristic that is likely to point towards wild disjunction, namely *non-localizability*:

> A system is localizable if the functional decomposition of the system corresponds to its physical decomposition, and each property of the system can be identified with a single component or subsystem. […] If system properties cannot be identified with components, but are instead distributed spatially within the system, that system is non-localizable. […] Higher level properties that are non-localizable are likely to have wildly disjunctive descriptions at the level of their components, and such properties are more likely to be irreducible to components (Sawyer 2001, 578).

At this point, however, it is important to note that Sawyer does not explicitly provide any argument for why it is thought that non-decomposability and non-localizability are indicative of wild disjunction, and thereby emergence. It may, however, not be difficult to see why. Remember that emergentists want to stress the importance of interactions. Also note that if a higher-level property is multiply realizable in a wildly disjunctive way, this means that what is of primary importance for its appearance is not the properties of the components and their linear relations that constitute its supervenience base—otherwise, its supervenience base wouldn't be wildly disjunctive. Instead, what seems to be more important and distinctive of the occurrence of such higher-level properties is the existence of *dense non-linear interactions* between the given realizing components. And so, given that both non-decomposability and non-localizability point towards the existence of such complex interactions, we can now see why they also point to wild disjunction.

I believe Sawyer would not object to this explanation as he himself goes on to suggest a fourth and final characteristic of wildly disjunctive systems, namely *complexity of interactions*. "The above criteria of non-aggregativity, non-decomposability, and non-localizability" he notes, "are all defined in terms of complex systemic relations among components. Consequently, several emergence theorists have suggested that the complexity of each interaction among components may be another variable contributing to emergence […]. Baas (1994) suggested that emergence occurs when the interactions are non-linear" (Sawyer 2001, 579).

To sum up, then, what all of the above characteristics of wild disjunction, and thereby emergence, seem to suggest is that the (properties of) component parts and/or their *linear* interactions/relationships that constitute the realization base are both necessary and sufficient for the manifestation of reducible properties. In the case of emergent non-reducible properties, however, the (existence of some) realization base in only necessary for their manifestation. What is further required for such properties to arise is the existence of appropriate complex non-linear interactions amongst the component parts of some appropriate realization base. And, as we shall see in the next three paragraphs, such properties are emergent because the dense interactive processes they originate from depend on the function of, and can only be defined by appealing to higher-level emergent systems, i.e., systems whose boundaries and properties transcend the boundaries and properties of their lower level realizing component parts.[105]

[105] We may now be able to see how this analysis may explain away Kim's explanatory causal exclusion problem. Remember, according to Kim, if there is an instance of causal relation between the physical properties of two events, then claiming that some higher-level property (that supervenes on the physical properties of its realization base) is also causally responsible for the occurrence of the events will be redundant; specifically, the higher-level property will be epiphenomenal. For instance, if C-fibers firing causally leads to A-chords vibration, then there is no need to claim that pain also led to A-chords vibration (i.e., screaming). In other words, if in order to explain one particular instance of screaming one can point to physical properties alone, then there is no need in positing (higher-level) mental properties as well. Here is a way out of this problem, however. Remember that pain might be multiply realizable in a wildly disjunctive way. That is, the physical properties and linear relations of the realizing component parts of every instance of pain may have nothing in common. Therefore, even though we may have a mere (as opposed to lawful) causal relation between C-fibers and A-chords, in order to have a *lawful* causal relation between pain and screaming in general (i.e., for every instance), we need to appeal to commonalities at a higher level of description. So, we need to appeal to properties that arise out of dense interactions between the realizing component parts, and which will (hopefully) be common for every instance. Now, such properties are emergent in that the dense interactions they arise out of can only be defined by appealing to higher level systems whose properties and boundaries transcend those of the realizing component parts. (Accordingly, pain may not be identical to the physical/chemical properties of C-fibers, because it may also depend on how C/D/E-fibers densely interact with other parts of the cognitive system. Pain, one might say in other words, may be identified with the functional role that C/F/E-fibers play in the larger cognitive system and not with the physical/chemical properties of C/F/E-fibers). Crucially, however, this is not to deny that higher-level mental properties are physical properties as well. It is only to say that they belong to a more complex level of physicality that cannot be captured by appealing to the lower-level properties of the component parts and their linear relations. Now, in this sense, even though mental properties inhere in the properties of their component parts and their linear relations (not all physical properties in any arrangement can give rise to mental ones), they, too, can be the cause of events. In particular, in order for the relevant events to occur (e.g., screaming) the properties of the parts of some realization base and their linear interactions must be in place. But what is further necessary and jointly sufficient with the former is the existence of the non-linear interactions between component parts, which the higher-level properties are identified with. So, higher-level mental properties and lower level physical properties are necessary and jointly sufficient for the occurrence of some physical event like screaming. (Although, notice, the lower level physical properties won't be necessary in the same way the mental properties will be, as the former can be multiply realized. That is, the existence of one of the appropriate realization bases will be necessary, but since there can be a multitude of them, no specific realization base is necessary. In contrast, the existence of very specific dense interactions and, thus mental properties, will be necessary).

Kim may object to this that in so arguing we have gone against the principle of the causal closure of the physical (i.e., all physical effects must have sufficient physical causes). Yet this is not a good objection since on this analysis mental properties are themselves physical—though physical at a higher level of complexity. The fact that these properties are

Let me explain further by first noting that the importance of such dense and complex non-linear interactions between components of the realization base for the emergence of systemic properties brings to mind some familiar considerations concerning dynamical systems theory. In section 2.3.2, we saw that when two components, or systems are mutually (and thus non-linearly) interdependent on the basis of feedback loops—i.e., when continuous reciprocal causation is manifested between components—there emerge coupled systems with new properties that are not possessed by any of their component parts. We further claimed that this is so on the basis of two arguments.

The 'ongoing feedback loops' argument went like this: when the effects of component A on component B are partly defined by component B's ongoing activity *at that time*, and vice versa, we cannot properly disentangle the two components in terms of distinct inputs and outputs from the one to the other. This is because the effects of each component to the other are not entirely intrinsic to the component itself, but they also depend on the affected component's ongoing activity. So the two units form a *causal amalgam*, which cannot be properly disentangled—not even in theory. We may now add that what this means is that the behavior of every component that forms part of this causal amalgam is no more entirely identical to its intrinsic properties, as it also hinges on (and has a further effect on) the behavior and, thus, on some of the properties of the other component.[106] Accordingly, such mutually determining and determined behavior will belong to both components at the same time and will disallow us to tell some of their properties apart. These will be (higher-level) properties of an overall emergent system that comprises of both subsystems that engage in the mutual interaction. Now, notice how

called 'mental' has only to do with the fact that they are associated with the mind. But, there is absolutely no necessity to further associate mind-talk with either substance, or property dualist considerations. Mental properties just refer to physical properties within a higher-level (physicalist) science of psychology.

(To generalize, then, according to this picture, higher-level sciences deal with the properties and causal powers of natural kinds that the natural kinds of the lower level sciences cannot capture. And the reason why they cannot capture them is because such higher-level kinds are multiply realizable, and thus their identification as kinds with specific properties and causal powers depends on a level of complexity that cannot be accounted for simply by referring to the properties and the linear relations of the kinds the lower-level science is appealing to).

[106] In what follows, I identify a component or system's properties with its behavior. According to DST, a system's dynamical law governs its behaviors, which can be represented as the system's flow on the system's phase portrait. The properties of the system are the geometrical properties of its flow.

this reasoning is very close, if not the same as, to the appeal to non-decomposability as a characteristic of emergent properties. In non-decomposable systems, the behavior of each component is directly dependent on the other components' behavior, and so it is no more intrinsically determined; it is impossible to determine all of the components' properties in isolation from the other components' properties. Accordingly, it will be impossible to draw firm boundaries between components. If we still want to point to an intrinsically determined system, such that all its properties will be determined in isolation from any other factors, we have to point to a larger emergent system that comprises of all subcomponents. In other words, contrary to the (lower-level) properties of component parts that can both be determined in isolation from the other components' behavior, and clearly individuate components, some of the properties of non-decomposable causal amalgams are interdependent properties that arise out of the component parts' mutual interactions. Such interdependent (higher-level) properties will not pick out any distinct parts but will instead be properties of the emergent non-decomposable causal amalgam as a whole.

Similarly, the 'systemic properties' argument held that when mutual interactions are in place, they give rise to new systemic properties that do not belong to any of the subsystems alone, but to the ongoing processes of interaction, which are internal to the overall coupled system. In some more detail, emergent properties are primarily the product of processes of dense non-linear interactions between system components. Those processes of interaction, however, cannot be identified with any component boundaries, but instead with the components' mutual interactivity. Accordingly, these properties belong to a larger system, comprising of all the components contributing to the relevant interactive processes. Moreover, since system individuation does not depend on any physical boundaries, but, instead, on the processes (and the properties they give rise to) one is interested in, and which emerge out of component interactions, the ontological postulation of coupled emergent systems seems to be necessary. Again, notice how this reasoning is very close to the appeal to non-localizability as a characteristic of emergent properties. In non-localizable systems, some properties cannot be identified with components, but are instead distributed spatially within the system. Accordingly, some of the properties of such systems won't belong to any subunit in particular, but instead to the overall system as a whole. In other words, such properties cannot be attributed to any component parts, since they do not arise aggregatively out of any (lower level) component properties and their linear relations. Instead, they emerge out of complex component interactions, which can only be defined by appealing to the overall emergent system.[107]

Now, before closing this section, here is what this second argument further demonstrates. Since, in system individuation, processes take priority over the physical boundaries of components (see section 2.3.2), and since higher-level systemic properties emerge out of processes of interactions, which define and are internal only to still wider systems, it appears that the existence of higher-level properties is tied to the existence of higher-level systems (i.e., entities). To be more precise, since one is primarily interested in processes (and/because of the properties they give rise to), and since mutually interactive processes (distinctive of emergent properties) can only be properly understood in terms of coupled and larger systems they belong to, the ontological postulation of such larger, emergent systems seems to be necessary. Contrary to Sawyer, therefore, we should not think that even though there could be emergent social properties, no social entities exist.

Apart from this last point, however, the upshot of the last paragraphs is that as in the case of extended cognitive systems, so in the case of distributed cognitive systems the existence of continuous reciprocal causation

[107] So let me try to be clear about the connection I try to establish between Sawyer's arguments and dynamical systems theory. Sawyer argues that a wildly disjunctive realization basis of a property points to the fact that the relevant property will be emergent and thus belong to an irreducible higher-level system of analysis. Here is the reason according to DST. A wildly disjunctive realization basis indicates that the relevant property does not really depend on the lower level properties and the linear relations of the components of its realization basis (as it can be wildly realized). Instead, it depends on appropriate dense interactions between the component parts of some appropriate realization basis. Due to two arguments from DST, however, such properties will be emergent. Here are the two arguments in brief: 1) When two lower level components mutually interact, we cannot individuate the interacting lower level components by telling their inputs/outputs, or properties apart, so *we cannot but appeal* to higher level systems that comprise of both of them and that are not recognized by the lower level analysis (this is the idea of non-decomposability). 2) *We must appeal* to higher level entities since the mutual interactions, and the properties these interactions give rise to depend on both systems at the same time, and thus to an overall system comprising of both of them, and which, again, is not recognized by the lower level system of analysis (this is the idea of non-localizability).

amongst their component parts is indicative of their emergence, and the emergence of their distinctive (higher-level) properties.[108]

5.3) Further Support for CRC as a Condition on Emergence

The upshot of the previous section, therefore, is the following. Even though the role of interactions has always seemed to be central in cases of emergent properties, it is not any kind of interactions that will do. Instead, it appears that only mutual, dense, non-linear interactions, which are indicative of continuous reciprocal causation between component parts can give rise to emergent properties and entities. The role of the present section will thus be to offer some independent theoretical support for the above claim, originating from research on distributed/group cognition. This will also allow us to become a bit clearer about how those interactions can give rise to distributed

[108] Let me finally offer one last example that will help us understand how non-reducible emergent properties and systems are supposed to appear on the basis of complex interactions between the component parts of their realization bases. So suppose we want to provide a lawful causal explanation of how agents may perform action X solely on the basis of chemical/physical terms. That is, suppose we want to reduce the occurrence of action X in the language of physics and chemistry. Notice, however, that agents might be multiply realizable in a wildly disjunctive way. So an agent's realization base may comprise of a biological perceptual system, a biological brain, a heart, and biological body parts, *or* by an artificial perceptual system, a silicon-based computer, a fuel pump, and artificial body parts. Now both agents may perform the same action X, whose occurrence, in both instances, depends on their respective realization bases. Yet the two causal explanations that in each instance led to action X are going to be fundamentally different because of the wild differences in the two agents' realization bases. Accordingly, no lawful causal explanation that only appeals to chemical/ physical properties can be provided such that it can capture both cases. That is, there is nothing common in the physical/chemical properties of, and the linear relations between the component parts of these two realization bases that explains how both of these two aggregates of realizing component parts performed action X. Instead, if one wanted to find such commonalities, one could only hope to find them at a higher level of description, i.e., at the level of dense interactions between component parts. So, in order to provide a lawful causal explanation for such wildly multiple realizations one can only appeal to the dense non-linear interactions between the realizing component parts. In so doing, however, due to DST (and, in particular, due to the 'ongoing feedback loops' and 'systemic properties' arguments), one must appeal to the higher-level system, namely to the agent, and its higher-level mental properties (i.e., the dense non-linear interactions of its component parts). And in so doing one must be a realist about this higher level of analysis not only because such entities and properties are irreducible to the entities and properties of the lower level of analysis (due to wild disjunction), but also because they do serious explanatory work. That is, the existence of some (but, not specific, due to multiple realizability) lower-level properties is necessary for the higher-level properties to be manifested, but in order for X to occur the higher-level properties (i.e., the dense interactions between the component parts) must also be in place. In other words, in order for X to occur, both higher and lower level properties must be in place. Adding a body, a heart, a brain, and perceptual system together in the right way without having them densely interact would never bring about X (and similarly for the artificial counterpart). Again, however, these dense interactions (i.e., emergent properties) derive from, belong to, and can only be accounted for by the higher-level system, i.e., the agent, which cannot be captured by the lower system of analysis (i.e., by lower-level properties of the lower level entities and their linear relations).

cognition and why this phenomenon in particular may not be reducible to the cognitive states/processes of individuals.

To start with, then, let us quote in length Heylighen at al. (2007) who invite us to

> consider a group of initially autonomous actors, actants or agents, where the agents can be human, animal, social or artificial. Agents by definition perform *actions*. Through their shared environment the action of the one will in general affect the other. Therefore, agents in proximity are likely to *interact*, meaning that the changes of state of the one causally affect the changes of state of the other. These causal dependencies imply that the agents collectively form a *dynamical system*, evolving under the impulse of individual actions, their indirect effects as they are propagated to other agents, and changes in the environment. It is important to note that a dynamical system has computational structure and is therefore able to process information. Not only that, but the dynamics themselves will generate a pattern, not just seek to complete it (Crutchfield 1998). Moreover, this system will typically be non-linear, since causal influences propagate in cycles, forming a complex of positive and negative feedback loops (6).

We thus see that Heylighen et al. also embrace the dynamical systems perspective in understanding group cognition, while also stressing the importance of mutual, dense and non-linear interactions amongst the individuals as a sign of the non-reducibility of their collective action. They then go on to note that all dynamical systems tend to self-organize. That is, they tend to evolve to a relatively stable configuration of states, i.e., what we termed in section 2.3.2 as an attractor of the dynamics. Once the system has achieved this stable configuration we can say that its component parts (in our case the agents) have mutually adapted by restricting their interactions to those that allow the collective configuration to survive. This process of "self-organization and further evolution of the collective effectively creates a form of "social" *organization*, in which agents help each other so as to maximize the collective benefit" (*ibid.*).

Furthermore, by taking a close look at this synergetic organization, they explain, we can distinguish between two fundamental dependencies between the agents' activities, or, more generally, the system's processes. First, two processes can use the same resource (input) and/or contribute to the same task or goal (output). Second, one process can be prerequisite for the next process (output of the first is input of the second). The first kind of dependence between processes or activities gives rise to tasks to be performed in parallel, and the second to tasks to be performed in sequence. The delegation of the right activities to the right agents at the right time is crucial for the efficient organization of the system, and this determines its overall shape. The parallel distribution of tasks determines the *division of labor* between the agents, whereas the sequential distribution determines their *workflow*. (Heylighen et al. 2007, 7).

Now, having these organizational considerations in mind, Heylighen et al. further note:

> Division of labor reinforces the specialization of agents, allowing each of them to develop an expertise that the others do not have (Gaines 1994; Martens 2004). This enables the collective to overcome individual cognitive limitations, accumulating a much larger amount of knowledge than any single agent might. Workflow allows information to be propagated and processed sequentially, so that it can be refined at each stage of the process. Self-organization thus potentially produces emergent cognitive capabilities that do not exist at the individual level. Moreover, it may give rise to unexpected organizational properties such as the emergence of a requirement of a new function, the loss of crucial information, the development of additional tasks and the deviation from existing workflow rules (Hutchins 1995) (Heylighen et al. 2007, 7).

So, as the authors note "self organization in this sense can be seen as the more efficient, synergetic use of interactions". However, they are fast to notice that the whole system is not entirely comprised by individuals. Interactions between agents necessarily pass through their shared physical environment, and we can call all the external phenomena that support these interactions *media*. Moreover, "certain aspects of the environment better lend themselves to synergetic interaction than others do. For example, a low bandwidth communication channel that is difficult to control and that produces a lot of errors, such as smoke signals, will support less synergetic interactions than a reliable, high bandwidth one, such as optical cable. Thus, there is a selective pressure for agents to preferentially learn to use the more efficient media, i.e. the ones through which causal influences—and therefore information—are transmitted most reliably and accurately" (ibid.).

They then go on to add that

> simply by using them, the agents will change the media, generally adapting them to better suit their purposes. For example, animals or people that regularly travel over an irregular terrain between different target locations (such as food reserves, water holes or dwellings) will by that activity erode

> paths or trails in the terrain that facilitate further movement. The paths created by certain agents will attract and steer the actions of other agents, thus providing a shared coordination mechanism that lets the agents communicate indirectly. A slightly more sophisticated version of this mechanism are the traits of pheromones laid by ants to guide other members of their colony to the various food resources […] Humans as specialized tool builders, excel in this adaptation of the environment to their needs, and especially in the use of physical supports such as paper, electromagnetic waves or electronic hardware to store, transmit and process information (*ibid.* 7-8).

We thus see that by this process of altering the media by using them, which in turn alters the way agents may interact and so on, "external media are increasingly assimilated or co-opted into the social organization, making the organization's functioning even more dependent on them. As a result, the cognitive system is extended into the physical environment and can no longer be separated by it" (Heylighen at al. 2007, 8).

Now, to pass to a different—yet substantially similar—defense of group agents, Theiner at al. (2010) hold that, on the basis of several case studies they explore, "specific cognitive capacities that are commonly ascribed to individuals are also aptly ascribed at the level of groups. These case studies show how dense interactions among people within a group lead to both similarity-inducing and differentiating dynamics that affect the group's ability to solve problems. This supports our claim that groups have organization-dependent cognitive capacities that go beyond the simple aggregation of the cognitive capacities of individuals" (378). Furthermore, they go on to explain, the underlying idea of their methodology is that the groups they are interested in have a particular structure, which is important to their behavior, including their ability to adapt to different circumstances. It is this organizational structure that allows them to speak of mechanisms of group cognition, which is not "simply the unstructured aggregation of individual cognition, but the outcome of a division of cognitive labor among cognitive agents. Such a division of cognitive labor may be the result of explicit organizational decisions by the individual agents, or (and we believe more commonly) the result of interactions among the agents that lead to enhanced group capacities without the express intent of the agents" (379).

Now, putting aside the empirical cases of group cognition the authors invoke—we will return to one of them (*viz.*, transactive memory systems) in the chapter to follow—Theiner et al. provide three reasons for which they think there are cognitive phenomena that require an explanation that goes beyond individual cognition. First, they claim, groups' behaviors depend critically on environmentally-enabled interactions between people. Environmental variables are part of the larger context in which people are situated, and are not reducible to individual cognition. We thus see that Theiner et al. are in complete agreement with Heylighen et al.'s point that the environment in the form of media is part of the overall distributed cognitive system. Second, Theiner et al. note that groups' solutions to problems may never be entertained by any individual, and in fact, the individuals may not even know that this is a problem they should be trying to solve. This, again, seems to be close to Heylighen et al.'s point that groups, which form dynamical systems tend to self-organize (that is, engage in the right sort of interactions), such that they, *as a group*, can operate competently and, thereby, survive. Finally, a third reason for which certain cognitive capacities can only be explained at the group level is that "the groups' solutions are not simply composites, unions, or intersections of individuals' solutions" (Theiner at al. 2010, 390). Once again, this brings to mind Heylighen et al.'s point that self-organization potentially produces emergent cognitive capabilities that do not exist at the individual level and which may give rise to unexpected organizational properties.

We thus see that there appears to be a consensus on the intuitions that may drive to collectivist considerations. But before closing this section, it should also be interesting to quote Theiner et al. when they explain what is required for a group of people to *constitute* a cognitive system in its own right. They write:

> For a group of two, or more people to constitute a cognitive system in its own right, we require that these people are coupled (in their functioning as members of the group, collectively performing a cognitive task) so as to form an *integrated system* with *functional gains*. Following Wilson (in press p. 19), we can break down the composite notion of a *functionally gainful, integrative coupling* as follows. First, two (or more) elements are coupled just in case they exchange information by means of reliable, two-way causal connections between them. Individuals who are collectively coupled are interdependent in their cognitive and behavioral activities. Second, two (or more) coupled elements form an *integrated system* in situations in which they operate as a single causal whole within the causal nexus—with causes affecting the resultant system as a whole, and the activities of that system as a whole producing certain effects. Third, an integratively coupled system shows *functional gain* just when it either enhances the existing functions of its

> coupled elements, or manifests novel functions relative to those possessed by any of its members. Distinguishing these two aspects of functional gain is important for an account of group cognition, because it implies that the cognitive interdependence between people has both individual-level as well as group-level effects (390-1).

As far as I can see, the above quote nicely summarizes the arguments and considerations originating from DST and which were presented in chapter 2 in favor of extended cognitive systems (and thereby extended cognition). These very same coupling arguments, we now see, can also be invoked in support of distributed cognition. Ongoing mutual interactions (i.e., continuous reciprocal causation) between agents and their artifacts can give rise to emergent processes (and properties), which belong to larger (cognitive) systems. Those are integrated systems in that the effects of each component directly affect the other components' activity, thus creating a causal nexus (or, a causal amalgam), where causes affect the systems as a whole. This integrated system can then give rise to new and unexpected properties both at the individual and the group level, and which may be mathematically perceived as new regularities in the coupled dynamical system's phase portrait (for example new attractors may appear and old ones may disappear). Those are qualitative changes, which can only emerge out of dense interactions between the system's subcomponents, and which may be properly ascribed only at the system as a whole.

5.4) Conclusion

To sum up, the present chapter focused on the metaphysical support of group cognitive agents and their properties. Accordingly, we came across two argumentative lines to this effect. On the standard debate between reductionism and non-reductionism, the existence of the phenomenon of wild disjunction provides a strong case for emergence (when wild disjunction is present). The same debate, reconstructed in mathematical terms, demonstrates that the coupling of two or more elements provides a clear verdict for emergent properties and entities. Even though seemingly distinct, those two arguments are far from unrelated. The link between these two approaches is that dense, non-linear interactions between component parts is indicative both of wild-disjunction and coupled/extended, or distributed cognitive systems.

We may, therefore, conclude that the existence of continuous reciprocal causation between individuals and the artifacts they use can give rise to group agents, whose function cannot be fully reduced to the individual level. Accordingly, we can consider such complex interactions as a necessary condition on the emergence of group cognitive agents.[109] Furthermore, those interactions give rise to synergetic behavior on the part of the group as a whole, which self-organizes into an appropriate structure such that it may function competently and thus survive (or more generally, fulfill its goal). Moreover, given the properties of the group members and the group's goal, this process of self-organization, which is the result of complex interactions between the members of the group and their artifacts, determines the division of labor and workflow of the whole structure, which may be the result of explicit organizational decisions, or not.

On the whole, we now have some considerable metaphysical support for the idea that there could be group processes, which give rise to group properties that may not belong to any individual alone, but to the group as a whole. It is time then to see how these considerations may apply to epistemology and the idea of *epistemic* group agents. More in particular, the fact that we now have a strong case for group cognition is an indication that we may also be able to make a strong case for robust anti-individualism in epistemology.

[109] It may be tempting to consider them as sufficient for group emergence as well. However, individuals could plausibly engage in such interactions while fighting, or being in war. Surely group entities could be identified in such cases, but not all individuals could belong at the same group, or system. Adding Clark's glue and trust criteria, however, could accomodate such counterexamples.

CHAPTER 6

Robust Epistemic Anti-Individualism

6.1) Introduction

The central aims of this chapter are to argue for epistemic robust anti-individualism, and understand the phenomenon through the lens of $\text{COGA}_{\text{weak}}$. I will start by providing two examples of epistemic group agents that satisfy the metaphysical criteria for group agency as revealed in chapter 5. Then, having these examples in mind I will ask what are the epistemic properties they have in common, and how can those common traits be accounted for by our preferred approach to knowledge, namely $\text{COGA}_{\text{weak}}$.

In some more detail, the previous chapter demonstrated that the idea of group agents can be motivated by the existence of non-reducible cognitive processes that emerge at the group level. What is required for such processes to emerge, I argued, is the manifestation of continuous reciprocal causation—on the basis of feedback loops—between the members of the group and their technological scaffolding, which may contribute both to the members' communication channels and their information-processing. Since, however, the focus here is on the cognitive task of knowledge production and acquisition, what we are here after is the existence of propositional knowledge that has been acquired by a collective reliable belief-forming process, which cannot be fully reduced to the set of cognitive abilities possessed by the individual members of the group—let alone the cognitive abilities of any individual in particular. If no such reduction is possible, then forming a true belief on the basis of the collective process won't be attributable to any particular member's cognitive agency either—not even merely to a significant, as opposed to a primary, degree. Accordingly, by the light of $\text{COGA}_{\text{weak}}$ two options will be available—reminiscent of Hardwig's (1985) two unpalatable conclusions—: either no one knows, or knowledge is not possessed by any individual alone.

Fortunately, however, being able to recognize groups as cognitive subjects in themselves, we can legitimately apply $\text{COGA}_{\text{weak}}$ at the group level, thus making sense of the claim that *p* is known by *S* even though it is not known by any individual alone (thus embracing the second of the two options above, and which amounts to epistemic robust anti-individualism). By way of giving flesh to this picture I will examine two cases: Transactive Memory Systems as studied by Wegner at al. (1985), and the case of knowledge acquired by research teams within scientific laboratories. Both cases demonstrate knowledge acquired at the group level, because in both cases true beliefs are acquired on the basis of reliable collective/distributed belief-forming processes that arise out of CRC between the members of the group and (if applicable) their artifacts as well.

What is more, these two examples of epistemic group agents have several properties in common that will help us understand how we can apply $\text{COGA}_{\text{weak}}$ to them. More specifically, in order to apply $\text{COGA}_{\text{weak}}$ at the group level we need to understand who is *S* in '*S* knows that *p*', what is the 'reliable belief-forming process' that gives rise to *S*'s true believing, what is *S*'s 'cognitive character' in which the collective belief-forming process must have been 'appropriately integrated', and, finally, what is the group's 'cognitive agency', to which the 'cognitive success must be significantly creditable', such that *S* can know that *p*.

6.2) Transactive Memory Systems

It should be rather uncontroversial to claim that memory is a reliable belief-forming process. Memory, that is, has a high propensity to deliver true rather than false beliefs. This is mainly the reason why memory is very often cited as a justificatory process on the basis of which an individual can claim to know some proposition *p* that he has encountered in the past. Interestingly, however, literature originating from cognitive psychology suggests that memory may be instantiated by more than just one individual on the basis of transactive memory processes. Accordingly, the first candidate for an epistemic group agent—that is a collective agent who can come to know on the basis of a belief-forming process that is not fully reducible to the cognitive abilities possessed by its individual members—will be a Transactive Memory System (TMS), (i.e., a group of individuals who collaboratively encode, store and retrieve information).

The reason why TMSs are good candidates for group minds—and thereby for epistemic group agents—is because, as Sutton et al. observe (2007), such systems are likely to involve skillful interactive simultaneous coordination of people, thus manifesting continuous reciprocal causation between individuals. As Wegner et al. (1985) claim, we can use TMSs in order "to conceptualize how people in close relationships may depend on each other for acquiring, remembering, and generating knowledge" (253). And they go on to further explain that it is primarily this 'cognitive interdependence' that motivated them to speak about group minds; when people come together, "they think about things in ways they would not alone" (254).

Now, before moving on to the details of a TMS, let us first take a look at two important methodological points that Wegner et al. stress. First, identifying a TMS as a group mind with respect to the process of memory does not imply a fine-grained similarity between the individual mental processes and the group mental processes. Instead, they are interested in the "*functional equivalence* of individual and transactive memory" (256):[110]

> Ordinarily, psychologists think of memory as an individual's store of knowledge, along with the processes whereby that knowledge is constructed, organized, and accessed. So, it is fair to say that we are studying "memory" when we are concerned with how knowledge gets into the person's mind, how it is arranged in the context of other knowledge when it gets there, and how it is retrieved for later use. At this broad level of definition, our conception of transactive memory is not much different from the notion of individual memory. With transactive memory we are concerned with how knowledge enters the dyad, is organized within it, and is made available for subsequent use by it (256).

Second, they explain, communication processes among group members are the center of group thought and it is communication processes that "*produce* the distinction between the group mind and the minds of individual members" (256). In some more detail, Wegner et al. hold that transactive memory systems are made up of two components: 1) an organized body of knowledge that is wholly contained in the memory systems of the individual

[110] Notice that proponents of extended cognition stress the exact same methodological point when they claim that there could be extended cognitive abilities that seem to be same in kind as organismic cognitive abilities (think, for example, about Otto's extended memory system). See also section 2.2.

members and 2) a repertoire of knowledge-relevant transactive processes that occur among group members on the basis of communication:

> Stated more colloquially, we envision transactive memory to be a combination of individual minds and the communication among them. This definition recognizes explicitly that transactive memory must be understood as a name for the interplay of knowledge, and that this interplay, no matter how complex, is always capable of being analyzed in terms of communicative events that have individual recipients. […] Using this line of interpretation, we recognize that the observable interaction between individuals entails not only the transfer of knowledge, but the construction of a knowledge-acquiring, knowledge-holding and knowledge-using system that is greater than the sum of its individual member systems (256).

So, having these preliminary considerations in mind, here is a first example of a transactive memory process:

> 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 that wonderful staffed 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 (257).

As Wegner at al. explain, during the discussion between Rudy and Lulu, the various ideas they exchange lead them through and elicit 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. And it is possible that without each other, neither Rudy nor Lulu could have produced the item" (1985, 257).

They go on, however, to note that transactive processes do not only take place during the retrieval of memories; they may, for example, occur during encoding as well. For instance, when partners perceive some event, each one may form some individual, private memory of it, but they may as well discuss about it along the way. Now, this discussion, far from being a mere rehash of the original event, could be much more. Here is an example of this transactive process of encoding memories:

> When a couple observes some event—say, a wedding—they may develop somewhat disparate initial encodings. Each will understand that it was indeed a wedding; but only one may encode the fact that the father of the

> bride left the reception in a huff; the other might notice instead the odd, cardboard-like flavor in the wedding cake. Their whispered chat during all this could lead them to infer that the bride's father was upset by the strange cake (Wegner at al. 1985, 259).

Since it is the group that generated the interpretation of the events, both partners will in the end encode the group's understanding of the events. In other words, their chat during the overall event will lead them to the encoding of a memory, which is qualitatively different from the memories they would have acquired, had they been on their own.[111]

We thus see that both encoding and retrieval may profit from transactive processes thus giving rise to group memories and knowledge that are qualitatively different from the memories and knowledge that one would have encoded and retrieved alone. In order for these transactions and their effects to occur, however, Wegner et al. note that certain conditions must be in place.

As they explain, "to build a transactive memory is to acquire a set of communication processes whereby two minds can work as one" (Wegner et al. 1985, 263). The first step in acquiring such a transactive memory system is to ensure that its candidate members will share a common culture and language so that they can adequately understand each other and, thus, communicate. In other words they must possess a common set of background assumptions. If this set of common knowledge is in place, the members of the group can begin a relationship, even as strangers, with a certain sense that each knows something that the other knows.

The second step is differentiation. Couples typically begin a relationship by revealing information about themselves to each other. Thus, in trading knowledge of their life goals, personality traits, emotional investments and so on, they are building the differentiation of their transactive memory; each fact about the self that is revealed to the other lends the other a sense of one's expertise and experience. As each member becomes more cognizant of the specialties of the other, the dyad's memory as a whole grows in differentiation.

[111] Notice that very similar transactive processes and effects may occur during the stage of memory storage as well. If, for instance, the dyad exchanges information not during the actual event but later on, when it is on its own, it may qualitatively alter the members' initial, individual memories of the event.

Now, this process of differentiation allows each member of the group to hold three kinds of information in his or her personal memory system. First, they may hold beliefs about the existence of higher-order information. Higher-order information can be thought of as the topic, theme, or gist of some set of lower-order information. For example, 'what George said' can be regarded as higher-order information for the actual words he spoke, which, in turn, constitute the lower-order information on this topic. Similarly, 'philosophy' is the higher order information for everything Orestis—a philosophy student—may know about this topic. Second, the above implies that each member of the group will also hold some lower-level information within their individual memory systems. And, third, they will also hold location information, which is information as to where any piece of higher-order, or lower-order information may be found. Now, communication in the dyad may transmit any of these three types of information. One may convey the lower-level information itself, or may just make known the existence of some higher-level information by saying "I spoke with George", or "I study philosophy". Moreover, in so doing, one immediately conveys location information, as one makes explicit that these facts are available in one's own memory. In Wegner et al.'s words then, "a differentiated structure, in this light, is one that contains mutual higher-order and location information, but reserves lower-order information for one or the other partner's memory alone" (264-65).

Now, once common knowledge and differentiation are in place, the dyad is ready for the final step towards the acquisition of a TMS, i.e., the formulation of an integrated structure. According to Wegner et al., an integrated structure comes about when members of a group combine their lower-level information about shared higher-order information so as to interactively produce knowledge of the higher-order topic that is qualitatively different from the lower-order information that they individually possess. Even though in all the previous examples the dyad manifested an integrated structure, here is yet another case:

> Imagine […] that a couple is leaving a party. At different times, they each talked to Tex. The male notes that Tex was depressed this evening; he stared at the floor and barely talked. The female says that Tex was not at all depressed; in fact, she saw him for quite a while early in the party and he seemed unusually frisky and friendly. The male recalls that Tex said he was

> thinking about separating from his wife. And in short order, the couple reached the conclusion: Tex was flirting with the female and feeling embarrassed about it in the presence of the male. (Wegner et al., 267)

So we see that integration of a TMS occurs when its members can productively take advantage of their differentiated structure so as to create new knowledge on the basis of effective communication feedback loops. In other words, once differentiation is in place, the dyad can move on to the integration process, whereby on the basis of the transactive processes of encoding, storage, or retrieval of information, it combines the lower-level information of its members in order to produce an integrated understanding of some higher-order topic. The existence of this integrated structure and its effects are the final indication that the group has successfully formed a TMS. Furthermore, these integrated bits of information—collaboratively produced on the basis of interactive feedback loops between the individual members—can, in turn, be stored on the individual memory systems of the members, thus becoming part of the common knowledge of the TMS, which can then be used in the future for even more effective transactive processes, and so on.

Now, having described TMSs in some detail it should be obvious that such systems manifest continuous reciprocal causation between their members at almost every dimension of their operation—i.e., during encoding, retrieval, storage, differentiation, and integration—thus clearly qualifying as a case of group mind/agent. Remember, in chapter 5, we argued that in order to have a group cognitive agent, its members must engage in continuous mutual interactions. The reason is that such interactive processes and the cognitive properties they give rise to, according to DST, can only be conceptualized by appealing to a larger system (i.e., the group agent) that comprises of all the individual members. Accordingly, such interdependent processes and properties do not belong to any individual alone; instead, they can only be ascribed to the group entity.[112] Moreover, given that memory

[112] Moreover, here is why such properties are not reducible to the properties possessed by the individual members, even though they supervene on them. Such properties are identified with the higher level interactions and not with the supervenient base that supports these interactions, because several supervenient bases that may have nothing in common at the individual level (and so resist lawful categorization on the basis of lower-level individual properties) could give rise to the relevant dense interactions. Therefore, the dense interactions, which are crucial to the group properties, cannot be identified with their actual supervenience base because in order to be manifested, they do not need any particular supervenient base, but only the existence of an appropriate one. Moreover, such properties

does, in general, count as a knowledge-conducive belief-forming process, and provided that the group's TMS is reliable, TMSs are good candidates for epistemic group agents (i.e., epistemic agents who come to know on the basis of collective belief-forming processes that are possessed only at the group level). So we now seem to have a first case of knowledge that is produced by a collective (memory) process that is not fully reducible to the cognitive processes possessed by the individual members of the group.

6.3) Scientific Research Teams

Inspired by Knorr-Cetina's (1999) ethnographic study of high energy physics (HEP) experiments in CERN, Hutchins' work (1995) on ship navigation, as well as Clark and Chalmers' (1998) hypothesis of extended cognition, Giere (Giere 2002*a*; 2002*b*; 2006; 2007; Giere & Moffat 2003) proposes to understand scientific experiments in terms of distributed cognitive systems. The main reason he offers for this is that, in much of scientific experimentation, "completing the task requires coordinated action by several different people" (2006, 711), and this coordinated action "makes possible the acquisition of knowledge that no single person, or a group of people without instruments, could possibly acquire" (2003, 305). He thus suggests understanding scientific experiments as emergent distributed cognitive systems that produce knowledge that no individual could on his/her own. And in order to understand this collective process of knowledge acquisition and production, that is, in order to "understand the workings of the big cognitive system one has to consider the human-machine interactions as well as the human-human interactions" (2002*b*, 292). The examples that Giere proposes are HEP experiments (2002*a*), the Hubble Space Telescope (2006), and Latour's (1999) (Giere & Mofat 2003) example of a scientific investigation that seeks to determine whether the Amazonian rainforest is encroaching on the adjacent savannah or if the savannah is encroaching on the rainforest.

Giere's focus on human-machine and human-human interactions should make clear that scientific experiments are good candidates for

will be properties of a higher system of analysis (here, the system of analysis that refers to group agents), because the dense interactions that are essential for their manifestation can only be conceptualized, according to DST, by a system of analysis that refers to systems whose boundaries and properties transcend those of the systems that the lower level system of analysis is able to deal with.

counting as epistemic group agents. Giere, however, goes nowhere close to providing a detailed description of the workings of the distributed cognitive systems he considers—no surprise, given their immense complexity—and neither will I. I will, however, provide a description of the interactive processes involved in an imaginary scientific experiment, which—even though an (over) simplification—captures the kind of complexity involved in real scientific investigations. The fact that real cases might be much more complex than my simplified scenario can then add to, rather than undermine, the argument. This added complexity and how it may further contribute to the overall discussion will become clear, later on, as we will also take a look at certain considerations from Knorr-Cetina's (1999) ethnographic study of HEP experiments.

The imaginary scientific experiment consists of a laboratory and three experts: an experimentalist, a mathematician, and a theoretical physicist. Before they enter the laboratory, of course, the three of them must come together in order to decide what instrumentation and analytical approaches they are going to employ. In other words, given their pre-experimental understanding of the relevant theoretical domain, and the available software and hardware tools (i.e., the available mathematical models and instrumentation, respectively), they must brainstorm in order to choose how they should tailor their scientific laboratory so as to most effectively study some proposition *p*, or a series of them.

Now, once agreement on these matters is achieved, the three experts can start the actual experimental work (for an illustration, consult the workflow, at page 154). The starting point of the experiment is the calibration of the instrument. In this first stage, the experimentalist runs several cycles of data collection in order to check whether they correspond with data in the literature, which are available from previous, similar experiments. In order to check the correspondence, of course, she will collaborate with the physicist who will interpret the available data for her. If the data turn out to correspond, the experimentalist will proceed to the next stage of collecting data relevant to the actual experiment. Otherwise, experimentalist and physicist will further collaborate in order to better calibrate the instrument. Once the instrument has been calibrated and sufficiently enough data have been collected, the experiment may proceed to its third stage.

This is the stage of data validation, whereby the mathematician confirms that the processed data are accurate and of good quality (i.e., there are no overlaps or weak peaks). If the results of data validation are not satisfactory, he will send them back to the experimentalist in order to perform a more accurate calibration, again, along with the physicist. If, however, the data are satisfactory, the mathematician will go on to analyze them. This is the process of identifying which theoretical entities are responsible for the observed signals in the sample. By its very nature, this stage cannot be performed by the mathematician alone; the contribution of the physicist is crucial at this point as well.

Now, once the data have been validated, and analyzed by both the physicist and the mathematician, the former can then study them on his own in order to come up with a model for them. During this stage, the physicist will use his knowledge and training in order to 'make the data fit' in mathematical models that are consistent with, or similar to models that have been previously used in the relevant theoretical domain—very rarely, but occasionally, he may even try to come up with an entirely new and surprising model. As soon, however, as the data have been satisfactorily modeled, the physicist will check his work with the mathematician. If the mathematician discovers any problems—say, with the calculations, or the arguments involved—he will send the model back to the physicist along with comments and suggestions, or the two of them may further collaborate in order to improve the model. Now, once the model passes satisfactorily the 'model check' stage, the physicist can work on an interpretation of it, thus, generating a working hypothesis with respect to the proposition(s) under consideration.

Once a working hypothesis has been generated the experiment can go through to its final stage, i.e., testing the hypothesis. At this stage, having the model at her hands—thereby knowing how the theoretical entities are supposed to behave with respect to varying conditions—the experimentalist can collect several data in order to check the validity of the working hypothesis. Should the data, after analysis and validation, match the predictions of the model—which will hardly ever happen from the first run of the whole experimental cycle—the three experts can finally come together in order to write an article and submit it for publication. Otherwise, the whole cycle begins all over again on the basis of more data that could be used in order to refine the proposed model and working hypothesis, or give rise to entirely new ones.

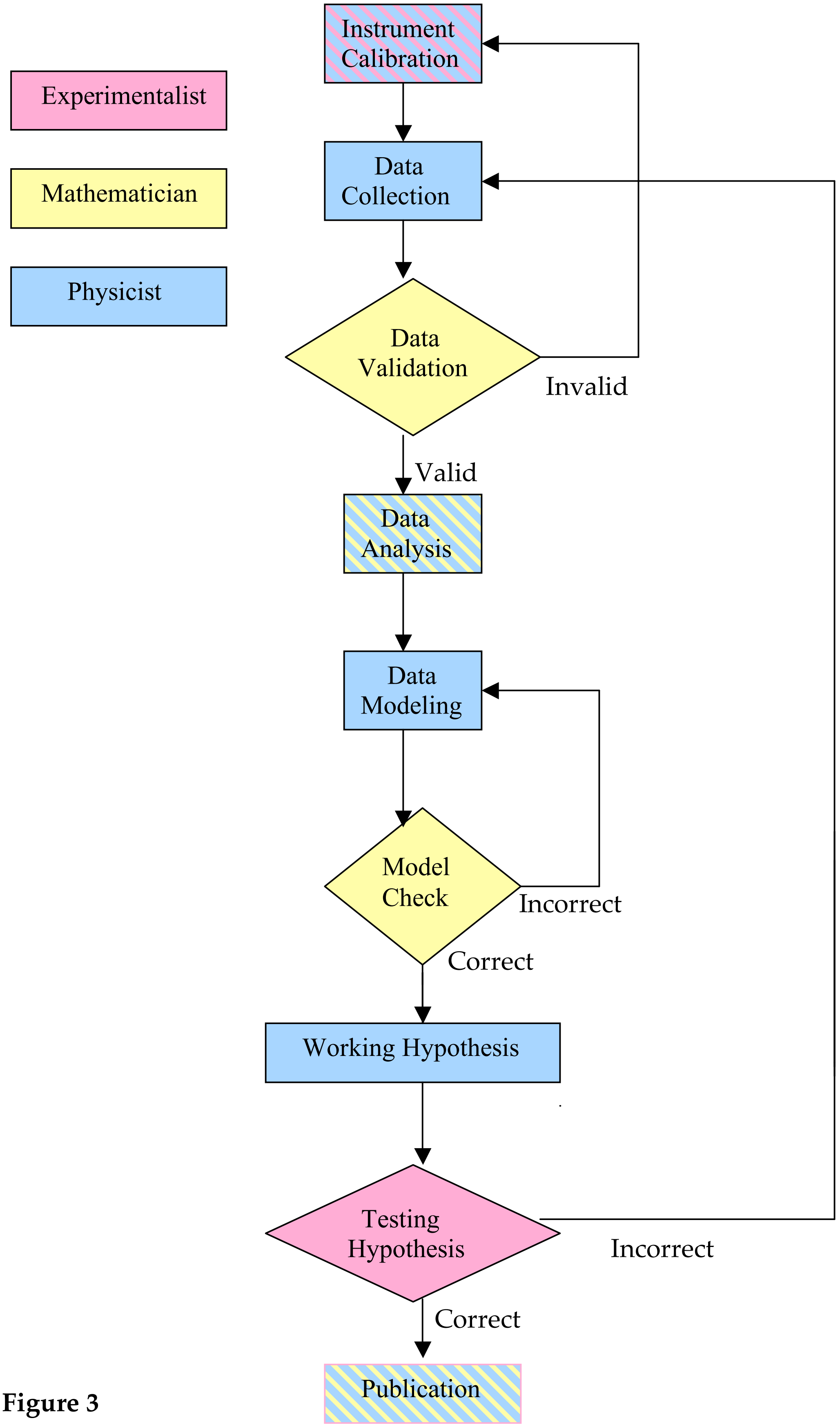

Figure 3

Now, to see why such experiments are good candidates for counting as group agents let me stress the following points. First notice that in order for the stages of instrument calibration, data analysis, and publication to complete, the direct collaboration of at least two experts is required (as indicated by the two- and three-colored boxes in the above workflow). Second, notice the direct feedback loops between the stages of 'data validation' and 'instrument calibration', 'testing hypothesis' and 'data collection', and finally, 'model check' and 'data modeling'. Now, add to these points the fact that—contrary to the first impression created by the workflow—not all of the above stages need to follow each other in discrete steps (for instance, the experimentalist may be continuously sending data to the mathematician for validation). As a result, the direct influence and continuous reciprocal causation between the activities of all experts should now be obvious. There is, then, a direct interdependence between these three individuals, which, as far as the process of experimentation is concerned, brings them together in a non-reducible emergent cognitive group. Put another way, as far as the processes of justifying and producing the final working hypothesis are concerned, they are not performed by any individual alone, but by the experiment as a whole, comprising of all the three experts and the instruments they use to collect and process data, and communicate between them.

Why should we include the instrumentation to the collective epistemic agent? First, as we saw in the previous chapter, Heylighen et al. (2007) explain that when a group uses some instrument in order to communicate or process information, altering the instrument by using it, which, in turn, alters the way the group members may interact and what they can do, "the external media are increasingly coopted into the social organization making the organization's functioning even more dependent on them. As a result, the cognitive system is extended into the physical environment and can no longer be separated by it" (8). In the previous example, this process of mutual dependence between the individuals and their instruments is most apparent during the instrument calibration, data collection, and data validation stages. Second, instruments are indispensable constitutive elements of the experiment in the more direct sense that during such experiments individual agents often extend their cognitive abilities (as discussed in chapter 2) to artifacts, by engaging in CRC with them. And since the epistemic group agent is constituted by the cognitive abilities of its individual members—which may extend to their instruments—the instruments are parts of the emergent collective entity as well.

Now, one possible worry that might be raised against the imaginary experiment presented above is that it is an oversimplification. Experiments are usually much more complex, involving more than just three experts and one experimental apparatus. It should be clear, however, that adding to the complexity of the experiment will actually generate more mutual interactions between scientists and their equipment, thus, reinforcing the idea that they form a collective entity. Is this, however, always the case? What about HEP experiments like UA2 and ATLAS, which have around 100, or even 2000 participants, respectively? Due to the amount of the participants involved, it may appear implausible that everyone, or at least most of the scientists taking part in such experiments will be able to directly influence each other.

In response, let us focus on what Knorr-Cetina says about HEP experiments in her book *Epistemic Cultures*, where she also argues for the "erasure of the individual as an epistemic subject" (166), on the basis of some "sort of distributed cognition" (173). Cetina explains that such large collaborations are not run by any individual alone and no individual is responsible for their management and organization. Instead, such experiments are managed by *discourse*, which "channels individual knowledge in the experiment, providing it with a sort of distributed cognition […], which flows from the astonishingly intricate webs of communication pathways" (173).

Cetina notes that this discourse may take place in numerous occasions, localities, and time slots. "Almost any situation, it seems, in which a participant finds him- or herself in the presence of another generates technical talk, including jogging in the vineyards and fields surrounding CERN or a bus ride to town" (174). Other opportunities for discourse may also occur in "the cafeteria", during lunch, drinks, and dinner, or around apparatus and equipment.

Apart from these informal cases, however, there is also a much more formal grid of discourse spaces created by intersections between participants, and "this grid was and is today perhaps the most important vehicle of experimental coordination and integration" (ibid. 174). More specifically, Cetina provides the following examples of the formally arranged discourse occasions that one may find in ATLAS: research and development meetings, panel meetings, institute meetings, steering group meetings, collaboration meetings, referee meetings, accelerator meetings, fixed committee meetings, special workshops dedicated to detector complexes, submeetings of some of the former, and the very important "meetings after meetings" (the informal exchanges that occur after scheduled events) (174).

Moreover, she notes, these meetings do not take place randomly, but they rather follow a specific sequence and temporal structure:

> The sequential structure and temporal structure of these meeting schedules is not without significance. As indicated, institute meetings placed at the end of several days of working group meetings and the plenary allow the earlier meetings to inform the institute meeting. The same succession of meetings existed in UA2, and it exists in ATLAS on the level of detector and R&D [(research and development)] collaborations. The sequential order suggests a passing of knowledge and technical decisions from the expert group where the responsibility lies to wider circles that take note of these details and play them back—through discussions, questions, and comments. Several rounds of this feed-forward and feedback result in the major technical decisions that are made by institute meetings (e.g., choices between competing detector technologies) after months of discussion. These rounds of discussion include panel feedback and panel recommendations, which are also channeled through working group meetings, submeetings, and plenary meetings (175).

Furthermore, Cetina accentuates the importance of the timing of those schedules in order for the above feedback and feed-forward communication loops to occur. "Schedules *pace, phrase,* and *state* the work, allocating turns within which certain points must be made or else points, and turns may be lost. For someone to hold their turn in the collective conversation, other activities (a study, a check, a calculation, an assembly task, a panel recommendation) must be performed on time and the results exhibited in status reports" (190). However, she also notes, schedules in HEP experiments are not in any individual's hands but they "originate from disperse and diverse sources, among which object requirements and the expertise of the researchers whom they regulate play a central role" (191). We therefore see that experiments, in Heylighen et al.'s (2007) terminology, 'self-organize' on the basis of schedules that derive from the workings of the experiment itself.

These schedules lock experimental activities firmly into deadlines and time slots, and they give rise to a 'common time ordering' of the individual activities, which constitutes a strong coordinating force that stitches the whole collaboration together.

As collaborations grow bigger and bigger, then, their participants do not lack the opportunity to directly interact with each other. To the contrary, as an ethnographic study of HEP experiments indicates, distant collaborators actually inform about and affect each other's work on the basis of formal avenues of discourse, ordered on a common time scale that coordinates everyone's work. As expected then, even in HEP experiments—at least in most of their aspects—most participants seem to engage in continuous reciprocal causation, which we have so far been using as an indication for the emergence of collective entities.

So, finally, in order to line up the discussion on scientific experiments with that on TMSs let me note that any collaborative scientific experiment must manifest the following characteristics. First, as in the case of TMSs, the participants of a scientific experiment must share some common knowledge on the basis of which they can communicate. Of course, contrary to TMSs this set of background knowledge does not have to be a common culture or language—as understood in the commonsensical way. Participants in such experiments may come from very disparate cultural backgrounds. Instead, in scientific experiments their members must share something close to what Kuhn termed as a paradigm: a set of metaphysical and methodological assumptions about what constitutes good scientific practice, as well as an agreement on and understanding of the broader set of theoretical tools and equipment they may employ. In other words, they must speak a common scientific language. Second, it must be clear enough that any scientific experiment, just like TMSs, will have a differentiated structure. That is, participants will hold beliefs about higher- and lower-order information as well as location information; participants will know each other's expertise and will know with whom they must communicate in order to retrieve higher- or lower-order information, which might be relevant to their work. An experiment will, thus, have a differentiated structure when its participants, just like in TMSs, have mutual higher-order and location information (i.e., information about everyone's expertise) but reserve lower-order information for one or the other expert and only retrieve it as required. And finally, given the aforementioned common knowledge and differentiated structure, a scientific collaboration will only count as integrated once its participants can take advantage of the above characteristics in order to interactively create new bits of knowledge which are qualitatively different from the knowledge each one could produce on one's own, and which they may later use to further progress their experiment.[113]

So, to move on to the next section, we have now seen two examples of epistemic group agents and their common characteristics. It is time then to see how $\text{COGA}_{\text{weak}}$ can account for knowledge possessed only at the group level (i.e., for robust anti-individualism in epistemology).

6.4) Epistemic Group Agents and $\text{COGA}_{\text{weak}}$

To recapitulate, in the previous two sections we came across two examples of groups of individuals that form a distributed/collective cognitive agent in virtue of the collective cognitive belief-forming processes that emerge out of the non-linear interactions between the members and the artifacts they use. Assuming, moreover, that their collective belief-forming processes are reliable we can claim that these are cases of epistemic group agents.

Let us then see how $\text{COGA}_{\text{weak}}$ can account for them. Here is $\text{COGA}_{\text{weak}}$ again:

[113] Here is a further characteristic of such collective entities concerning the members' communication, and which may have escaped our attention both in the discussion of TMSs and scientific experiments. In such cognitive collaborations, experts will communicate on a certain kind of trust, which is rather distinct form the kind of trust exemplified in ordinary cases of testimony. In discussing testimonial knowledge in chapter 4, we mentioned that in order for the hearer to accept the speaker's testimony he must check both that there are positive reasons for, and no negative ones against doing so (i.e., the hearer must check for the reliability of the report). Not so with communication within group cognitive agents. In the case of collective epistemic agents each expert will by default accept another expert's reports without checking for its reliability. This is probably because they have a common goal—the group's goal—and they also share a paradigm/common language, which means that they should, in principle, agree with and recognize as reliable the methods their informant used in order to arrive to the offered report, even though, in practice, they may know nothing about them (i.e., their methods).

In other words, one further characteristic of group cognitive entities is that each member is by default warranted in accepting every other member's reports. What this means is that the transmission of knowledge in such cases is not strictly testimonial, as it is knowledge that does not really originate from outside the epistemic agent but is instead circulated within it.

> If *S* knows that *p*, then *S*'s true belief is the product of a reliable belief-forming process, which is appropriately integrated within *S*'s cognitive character such that her cognitive success is to a significant degree creditable to her cognitive agency (Pritchard 2010*b*, 136-7).

So, to begin with, according to a robustly anti-individualistic reading of COGA$_{\text{weak}}$, who is *S*? As previously noted, we can individuate group agents on the basis of the collective processes their component parts give rise to. The reason for this is that such collective processes emerge out of subsystems interacting continuously and reciprocally, which behavior, according to DST, leads to the postulation of an overall system, because in its absence the relevant collective processes cannot be understood. Since, however, not only human agents but also artifacts may take part in the relevant interactions, the upshot is that the collective epistemic agent, *S*, may be comprised of both individuals and the artifacts they use in order to process information and communicate.

Let's now turn to the notions of a reliable belief-forming process and *S*'s cognitive character. In the discussion of virtue reliabilism in chapter 1, we noted that, according to Greco, one's cognitive character is the set of reliable and stable dispositions (i.e., belief-forming processes), as well as one's memories and overall doxastic system, that one manifests when one is conscientious (i.e., motivated to believe what is true). Now, there is no reason why we should understand an epistemic group agent's cognitive character any differently. Notice, however, that epistemic group agents usually come together to form just one collective belief-forming process—in virtue of which they also exist. We might assume, then, that, usually, an epistemic group agent's cognitive character will be identical to just that one collective belief-forming process. Greco, however, includes to one's cognitive character one's memories and overall doxastic system as well. Perhaps, then, we should also add within an epistemic group agent's cognitive character the shared common knowledge of its members (i.e., their paradigm in cases of research teams)—including their common knowledge of the group's differentiated structure—which allows for and provides the framework for their collective belief-forming process to arise, and, maybe, even make sense.

Here are also some thoughts regarding the reliability of the collective belief-forming process and its relation to a conscientious epistemic agent.

Remember that according to Heylighen et al. (2007) individuals form collective entities on the basis of the process of self-organization so as to bring about a desired result or maximize the collective benefit. That is, collections of people tend to interact until they evolve to a stable configuration of states. Once the system has achieved this stable configuration we can say that its component parts have mutually adapted by restricting their interactions to those that allow achieving their end (the end, amongst other things, could be fitness, profit, or, in our case, objectively reliable true beliefs). Now, given that we here focus on epistemic group agents who are also conscientious, i.e., motivated to believe what is true, it follows that their process of self-organization, which gives rise to the relevant collective belief-forming process, will also ensure the latter's reliability. Otherwise, the collective agent would have not accomplished its end, thereby dissolving, or would have given rise to another internal configuration and, thus, to a different belief-forming process that would have been more appropriate (i.e., objectively reliable). Notice, moreover, that as in the case of individual epistemic agents, in order for the group to be also subjectively justified, it does not need to have any beliefs regarding the reliability of the collective belief-forming process. Instead, provided 1) that the belief-forming process is a disposition of the group agent, and 2) that the group is motivated to believe what is true (such that it would not employ any belief-forming processes that have, in the past, notably failed in being conducive to this end) we can make the following claim. The group will be subjectively justified in holding the resultant beliefs merely by *not* having any beliefs that the collective belief-forming process has been notably problematic in the past. (In Clark's terms, the collective belief-forming process must not usually be subject to critical scrutiny).

So, let us finally turn to the idea of appropriate integration, which will also allow us to explain what could count as the epistemic group agent's cognitive agency. Back in section 3.2, where we were discussing when an external element can count as properly integrated within one's cognitive character, we claimed that in order for this to be the case, the external element must be continuously and reciprocally interacting with the agent's organismic cognitive faculties. That is, in order for an artifact to be a constitutive element of one's cognitive economy, the agent must deliver on its basis outputs which recycled as inputs will drive his cognitive character along. Therefore, in order for a belief-forming process to be appropriately integrated within one's cognitive character, the phenomenon of CRC must be manifested between the target process and one's organismic cognitive faculties. We also noted that the underlying reason for this is that HEC clearly acknowledges the central role of the persisting biological organism in order to eventually accomplish its very own cognizing: "Human cognitive processing (sometimes) extends to the environment surrounding the organism. But the organism (and within the organism the brain/CNS) remains the core and currently the most active element. Cognition is organism centered even when it is not organism bound" (Clark 2007, sec. 4). And this, we further noted, explains why using the CRC phenomenon to judge whether an artifact has been appropriately integrated within one's cognitive character is in line with COGA$_{\text{weak}}$, which demands that the cognitive success be, in the end, significantly creditable to one's cognitive agency. It is the agent's organismic cognitive faculties (i.e., the brain/CNS) that are first and foremost responsible for the appropriate employment and recruitment of the external elements, on whose basis the agent will deliver outputs, which recycled as inputs will drive her cognitive character further along, so as to eventually form a true belief with respect to some proposition *p*. (Notice here, then, that I, in effect, identify an agent's cognitive agency with the agent's organismic cognitive faculties).

An analogous understanding of the process of integration and of the notion of cognitive agency should, I think, be provided in the case of epistemic group agents. Notice, however, that the case of distributed cognition and group agency does not involve merely an artifact being employed by an agent, but is, instead, about many agents and their artifacts coming together to form a cognitive system comprising of all of them. So, it is not really apposite to speak of any agent, or artifact being integrated within anyone else's cognitive character; instead, we should rather ask when the whole group of individuals and their artifacts counts as having an integrated structure. In other words, we need to know when a potentially shared belief-forming process is going to count as having an integrated structure that can give rise to *S*'s shared cognitive character—and thereby to *S*. And this, according to what has been said about group agents (see chapter 5 and especially sections 5.2.3 and 5.3), won't happen unless *S*'s members engage in

CRC between them and their instruments—notice that this is very much in agreement to Wegner's view regarding the integration of TMSs.

Moreover, demanding CRC so as to have a shared integrated cognitive character also points out how we should understand a group's cognitive agency. Since CRC within a group is manifested and maintained primarily due to the organismic cognitive faculties of its members, it is the set of these organismic cognitive faculties that is first and foremost responsible for the emergence of distributed cognition, and which should count as the group's cognitive agency; to paraphrase Clark, cognition is organism centered even when it is distributed. So, an epistemic group agent's shared reliable belief-forming process—which may consist of both humans and instruments—is primarily the product of dense interactions between the organismic cognitive faculties of its members. Accordingly, we can see why—and in accordance to COGA$_{weak}$—the collective cognitive success must be significantly creditable to the group's cognitive agency (i.e., to repeat the claim, the set of the organismic cognitive faculties of its individual members). It is the set of these organismic cognitive faculties that is first and foremost responsible for the emergence and appropriate employment of the collective belief-forming process so as to eventually accomplish its own cognizing.

Crucially, furthermore, the above considerations also indicate why knowledge, in such cases, cannot be attributed to any individual alone, thus, giving rise to robust anti-individualism in epistemology. As noted in section 2.3.2, whenever components exhibit CRC between them, they give rise to an overall system consisting of all of them, without any single component causally standing out of the whole. In other words, the activity of the system depends on the system as a whole and it is to the system as a whole that the success of its activity must be attributed. Therefore, in cases of collective cognitive success that is the product of the interactions of the organismic cognitive faculties of the members of the group, the success must be attributed to the set of the members' cognitive agencies as a whole (i.e., to the cognitive agency of the group), and to none of the individual members alone. And, to anticipate our final conclusion, given that, according to COGA$_{weak}$, knowledge must be creditable and reliably formed true belief, such beliefs won't be known by any individual alone, but by the group agent as a whole, because it is only to the group agent's cognitive agency that the cognitive success can be significantly attributed.[114]

6.5) COGA$_{weak}$ and Robust Epistemic Anti-Individualism

Here is then the overall picture motivating robust anti-individualism in epistemology. There are cases where individual epistemic agents come together with the motivation to believe what is true. This leads them to interact as a group, which self-organizes so as to believe what is true. For the sake of truth and on the basis of interactive self-organization there emerges a reliable collective belief-forming process that constitutes the group's shared cognitive character, and in virtue of which we have a non-reducible epistemic group agent. The reason why this is a non-reducible epistemic group agent is that its cognitive character emerges out of dense, non-linear interactions between its individual members and their artifacts, and according to DST, such emergent properties or processes cannot be conceptualized in the absence of the system as a whole, which is more than the sum of its parts. Accordingly, any cognitive success achieved on the basis of a collective belief-forming process cannot be fully reduced to the sum of the cognitive abilities of the participating individuals. It can, however, be at least significantly attributed to the set of all the participating individual members of the group as a whole, as it is the set of their cognitive agencies that is first and foremost responsible for recruiting and maintaining the overall group agent.[115]

Due to its collaborative nature, therefore, this is cognitive success that is not produced by, and, thereby, not creditable to any individual alone. Should we then conclude that such collaboratively, socially produced beliefs cannot count as knowledge, or should we embrace Hardwig's second conclusion: there can be knowledge that is not known by any individual alone? What the above discussion indicates is that we can opt for this second

[114] Put another way, since the justificatory process is a collective one—arising out of the members' interaction—none of the members could have come to know the proposition on their own. This point can be further motivated, if we also consider the differentiated structure of group agents. Members contribute only differentially to the collective belief-forming process and so the cognitive success cannot be significantly attributed to any specific individual alone.

[115] Furthermore, if the group has not produced its own artifacts then, in the spirit of chapter 4, the rest of the credit should be attributed to the broader epistemic community that brought those artifacts about.

option; since the cognitive success of such collaboratively produced true beliefs can be significantly attributed to the group agent's cognitive agency, we can claim that such beliefs, according to COGA$_{weak}$, can be known by *S*, *the epistemic group agent*. In other words, being able to recognize group agents as epistemic (*qua* cognitive) subjects in themselves, we can now apply COGA$_{weak}$ to such robustly anti-individualistic cases as well. Therefore, we can now make sense of the claim that *p* is known by *S*, even though it is not known by any individual alone.

And finally, let me also note that this is not to claim that a piece of collectively produced propositional knowledge cannot be individually known. As soon as the group publishes its results, or in anyway, testifies them, one can come to know them individually. The initial justification of the proposition, however, will always lie with the group. So, if one asked to reproduce that initial justification, then, unless technology has progressed so much as to simulate the group agent, one would have to recreate the group structure that gave rise to the initial collective belief-forming process. Moreover, this possibility of gaining—on the basis of testimony—individual knowledge of a collectively produced true belief—which cannot be otherwise produced—appears to be a good case of downward causation within epistemology, whereby some non-reducible social entity affects the epistemic state of an individual.

AFTERWORD

I have here argued that COGA$_{weak}$—a virtue reliabilistic necessary condition on knowledge—allows to reveal and account for the social nature of several kinds of knowledge. More specifically, COGA$_{weak}$ can easily handle testimonial knowledge and knowledge of coverage supported beliefs. But reading COGA$_{weak}$ along the lines suggested by HEC can also explain how one may acquire knowledge on the basis of the operation of epistemic artifacts, while remaining fast to the ability intuition on knowledge. Now, to the point, all of these three cases reveal the dual nature of knowledge. That is, in these cases, knowledge is not only attributable to the individual agent that holds the true belief, but also to other individuals who contributed to the individual's reliable formation of her true belief either directly or indirectly. All three cases, therefore, point towards weak anti-individualism in epistemology. We also noted, however, that reading COGA$_{weak}$ through the spectacles of HDC can help us account for robust anti-individualism in epistemology, i.e., for knowledge that is not possessed by any individual alone. This is so, because there are certain cases whereby a true belief is not reliably formed by any individual's cognitive ability, but by a collective cognitive ability that belongs to a group agent. Accordingly, this kind of knowledge cannot be attributed to any individual alone, but to the group agent as a whole. Still, however, since on the basis of HDC we can recognize group agents as epistemic subjects in themselves we can legitimately apply COGA$_{weak}$ to such robustly anti-individualistic cases as well. In other words, we can now make sense not only of the claim that knowledge might be both social and individual, but also of the claim that a proposition might be known by *S* even if it is not known by any individual alone.

So the overall picture is this: using (active) externalist philosophy of mind to interpret one of the most popular mainstream individualistic approaches to the theory of knowledge (that of virtue reliabilism)—as captured by COGA$_{weak}$—allows to account even for the most provocative claims that one could come across within social epistemology (that of robust anti-individualism). In other words, externalist philosophy of mind is the means by which we can reconcile individualistic and social epistemology, while remaining fast to mainstream epistemology.

So, let us try to localize the present account within the general field of epistemology in some more detail. $\text{COGA}_{\text{weak}}$ is clearly an analysis of knowledge in individualistic terms. For this reason it fits well within the realm of mainstream contemporary epistemology. Pointing, however, to the social nature of many instances of knowledge, the present account also fits within the field of social epistemology. In which exact sense? In "Why Social Epistemology Is *Real* Epistemology" (2009) Goldman distinguishes social epistemology in two branches. The first, 'preservationist' social epistemology, preserves the core individualistic assumptions of mainstream epistemology while studying themes such as testimonial knowledge and peer disagreement. The second branch, 'expansionist' social epistemology, is supposed to distance itself from some of the individualistic assumptions of mainstream epistemology as it is mainly concerned with the topics of the epistemic properties of group doxastic agents, and the influence of social systems and their policies on epistemic outcomes.

Wherein does the present account best fit? We have seen how $\text{COGA}_{\text{weak}}$ accounts both for 'preservationist' topics such as testimonial knowledge, and 'expansionist' topics such as group doxastic agents. The answer, therefore, is that it fits both within the scope of preservationist and expansionist social epistemology. Crucially, however, in both cases, it does so by remaining fast to individualist mainstream epistemology. Therefore, one could claim, reading $\text{COGA}_{\text{weak}}$ through the spectacles of HEC and HDC provides a reconciliation of individualist epistemology with both preservationist and expansionist social epistemology.

What about the history and philosophy of science? After the second half of the 20th century, and in particular the current of Historicism as initiated by Kuhn's *The Structure of Scientific Revolutions*, hardly anyone could deny the social nature of the scientific process. This realization, however, seemed to pose a problem for the application of mainstream individualistic epistemology to the study of science. The present account, however, could now provide a useful link between the two. In particular, science is primarily performed by individual scientists employing their hardware and software epistemic artifacts, or by research *teams* operating within scientific labs, which are uniquely tailored to fit their purposes. Extended epistemic agents, and epistemic group agents could, therefore, become very handy for a mainstream epistemological analysis of the scientific progress. Indicatively, discussing the scientific revolution of the 16th century, Giere and Moffat (2003, 308) write:

> 'No "new man" suddenly emerged sometime in the sixteenth century….The idea that a more rational mind…emerged from darkness and chaos is too complicated a hypothesis' [Latour 1986, 1]. We agree completely. Appeals to cognitive architecture and capacities now studied in cognitive sciences are meant to explain how humans with normal human cognitive capacities manage to do modern science. One way, we suggest, is by constructing distributed cognitive systems that can be operated by humans possessing only the limited cognitive capacities they in fact possess.

And elsewhere (2002*b*, 298) on the same topic, Giere alone writes:

> It is often claimed that the scientific revolution introduced a new way of thinking about the world, but there is less agreement as to what constituted the 'new way'. The historiography of the scientific revolution has long included both theoretical and experimental bents. Those on the theoretical side emphasize the role of mathematics, Platonic idealization, and thought experiments. The experimentalists emphasize the role of experimental methods and new instruments such as the telescope and microscope. Everyone acknowledges, of course, that both theory and experiment were crucial, but these remain a happy conjunction.
>
> The concept of distributed cognition provides a unified way of understanding what was new in the way of thinking. It was the creation of new distributed cognitive systems. Cartesian co-ordinates and the calculus, for example, provided a wealth of new external representations that could be manipulated to good advantage. And the new instruments such as the telescope and microscope made possible the creation of extended cognitive systems for acquiring new empirical knowledge of the material world. From this perspective, what powered the scientific revolution was an explosion of new forms of distributed cognitive systems. There remains, of course, the historical question of how all these new forms of cognitive systems happened to come together when they did, but understanding the source of their power should now be much easier.

Now, as Giere himself notes in the end of the above quote, 16th century scientific revolution was probably also the result of further underlying cultural changes that may go unobserved, at first pass, by the historian and philosopher of science. But the point I want to make remains the same. Combining the frameworks of externalist philosophy of mind and externalist epistemology is a significant step forward towards a deeper understanding of the workings of the social epistemic project that is called science.

So let me finally close by mentioning some potential topics of future research that this thesis may hint to. In this essay, I have only used active externalism within philosophy of mind to demonstrate the partly or entirely social nature of certain kinds of knowledge. Since, however, as I explained in my foreword, active externalism (in the form of HEC and HDC) is the extreme consequent of the embodied and embedded paradigm shift within cognitive science, one could expect that these latter, less provocative theses could be of interest to epistemology as well. This, admittedly, could be the topic of a second thesis. In particular, since epistemic agents are both embodied and embedded, the embodied and embedded approach in cognitive science could resolve many issues both in individualist and social epistemology that I have here passed in silence. For instance, how can changes in one's body or environment affect the reliability of one's cognitive abilities? And, how can we manipulate such changes to good (epistemic) effect? How can we design an epistemic agent's environment, such that we can maximize her epistemic benefits? Or, in which specific ways one's society may affect one's epistemic standing? Could we tailor epistemic societies such that the problem of peer disagreement would be less worrying? What is it about contemporary society that allows for an exponential growth of knowledge and technology?

Moreover, even though I have argued that $\text{COGA}_{\text{weak}}$ can in general account for knowledge possessed at the group level, I have gone almost in no detail about the specific epistemic/cognitive properties that each time allow such entities to achieve their epistemic ends, and how such ends may inhibit or reinforce the (epistemic or practical) goals of their individual members. So in the future, and having the present account at hand, it should be interesting to see how one could maximize the epistemic benefit of such collective entities as a function of their organization, and how this maximization may stand in the way, or promote the needs of their individuals members. Such an in depth analysis of the epistemic properties of group agents and their members could profit from both a purely theoretical approach, and the study of several historical and contemporary case studies, originating from the fields of the history and philosophy of science.

And finally, turning back to the individual and given the combination of the hypothesis of extended cognition and $\text{COGA}_{\text{weak}}$, how can we further boost (i.e., extend) the epistemic agent’s cognitive capacities? In other words, how can we create epistemic artifacts that are more reliable, transparent in use, and easier to integrate within one’s cognitive character? Epistemology can explain both how knowledge is actually being achieved, and how it can be further achieved in new ways.

Glossary of Abbreviations and Technical Terms

Ability Intuition (on knowledge): Knowledge must be the product of cognitive ability.

Anti-luck Intuition (on knowledge): Knowledge must not be due to luck.

Attractors: Limits sets that gravitate trajectories passing through all nearby states. Attractors also govern the long-term behavior of a system.

Autonomous System: A system whose dynamical law depends only on the values of its state variables and the values of some set of fixed parameters.

Basin of Attraction: The set of initial states that converge to a given attractor.

Bifurcations: Qualitative differences in the system's flow as a result of changes in the system's parameter values.

CNS: Central nervous system.

COGA$_{\text{weak}}$: If *S* knows that *p*, then *S*'s true belief that *p* is the product of a reliable belief-forming process, which is appropriately integrated within *S*'s cognitive character such that her cognitive success is to a significant degree creditable to her cognitive agency. (Pritchard 2010*b*, 136-7)

Cognitive Agency (of individuals)**:** One's organismic cognitive faculties of the brain/central nervous system.

Cognitive Agency (of group agents): The set of organismic cognitive capacities of its individual members, as a whole.

Cognitive Character: The dispositional/habitual belief-forming processes that one manifests when one thinks conscientiously. It consists of one's cognitive faculties of the brain/central nervous system, including one's natural perceptual cognitive faculties, one's memories and the overall doxastic system. In addition, it can also consist of acquired habits of thought, "acquired skills of perception and acquired methods of inquiry, including those involving highly specialized training or even advanced technology" (Greco 1999, 287).

Coupled System: An autonomous system that consists of two (or more) nonautonomous systems that mutually interact in that the parameters of each system function as state variables of the other, and vice versa.

'Coupling-Constitution' Fallacy: *Simple Version*: "It simply does not follow from the fact that process X is in some way causally connected to a cognitive process that X is thereby part of that cognitive process" (Adams & Aizawa 2008, 91). *Systems Version*: It unfolds in two steps: "The first is to move from the observation of some sort of causal connection to the claim that the brain, body and relevant parts of the world form a cognitive system. The second step is a tacit shift from the hypothesis that something constitutes a system to the hypothesis that is an instance of extended cognition" (Adams & Aizawa 2008, 92). This, however, is again fallacious: "It simply does not follow from the fact that one has identified an *X* system in terms of a causal process of type *X* that that process pervades every component of the system" (*ibid*., 125).

CRC: Continuous Reciprocal Causation (between objects, components, systems and so on) .

DST: Dynamical Systems Theory.

Dynamical Law (of a system): A set of differential equations that regulate the change of the state variables of the system.

Epistemic Externalism: The denial of epistemic internalism.

Epistemic Internalism: One knows that *p* only if one has, in principle, internal access (i.e., by reflection alone) to the reasons for one's belief that *p*.

Externalism (in philosophy of mind): Cognition may extend beyond the agent's organismic boundaries to epistemic artifacts, or distributed across individuals and their epistemic artifacts.

Flow (of a system): The entire range of the possible trajectories of an abstract dynamical system.

HEC: Hypothesis of Extended Cognition.

HEMC: Hypothesis of Embedded Cognition.

HEP: High Energy Physics.

Input: The parameters of nonautonomous systems that vary across time.

Internalism (in philosophy of mind): Cognition is restricted within the agent's head, or, at most, her organismic boundaries.

Limit set: Sets of points on a system's state space that are unaffected by the dynamical law, in that if the state of the system enters a limit set, the dynamical law will keep it there indefinitely.

Nonautonomous System: A dynamical system in which, in addition to its state variables, one or more of its parameters vary across time, as well.

Parameter Space: A theoretical space with geometrical properties that includes the values of all the parameters encountered in a mathematical model.

Phase Portrait: A graphical representation of the different phases (sequences of trajectories) the system might enter into, given its attractors and repellors.

Process Reliabilism: Knowledge is the product of a reliable belief-forming process.

Repellors: Limit sets that are unstable in that some nearby trajectories diverge from them.

Robust Epistemic Anti-Individualism: Knowledge that is not known by any individual alone.

Robust Epistemic Individualism: Knowledge must be fully down to the individual.

State Space (of a system): A theoretical space with geometrical properties that includes all the possible values of the state variables of the system.

TMS: Transactive Memory System.

Trajectory: A sequence of states the system will enter, given its dynamical law and some initial state x_0.

Transients: The portions of trajectories that are found within a basin of attraction, but which do not lie in the attractor itself.

Vector Field: A field on a system's state space that determines the direction of the system's trajectories.

Virtue Reliabilism: *S* knows that *p* if and only if *S*'s reliable cognitive character is the most important necessary part of the total set of causal factors that give rise to *S*'s believing the truth regarding *p*.

Weak Epistemic Anti-Individualism: Knowledge is partly individual and partly social.

Weak Epistemic Individualism: Knowledge must be primarily down to the individual.

REFERENCES

Adams, F., and Aizawa, K. (2008). *The Bounds of Cognition,* Blackwell Publishing Ltd.

—— (2010). 'Defending the Bounds of Cognition'. In *The Extended Mind.* (2010), Menary (ed.) Cambridge, Massachusetts, MIT press.

—— (2001). 'The bounds of cognition'. *Philosophical Psychology,* Vol. 14, No.1, 43-64.

Andersen, B., P., Emmeche, C., Finnemann, N, and Voetmann, C., (eds.) (2000). *Downward Causation: Minds, Bodies and Matter.* Aarhus: Aarhus University Press.

Audi, R. (1998). *Epistemology: A Contemporary Introduction to the Theory of Knowledge.* London: Routlege.

Baas, N., A. (1994). 'Emergence, Hierarchies, and Hyperstructures'. Pp. 515-37 in *Artificial Life III*, edited by Christopher G. Langton. Reading, Mass.: Addison-Wesley.

Beer, R. (1995). 'A dynamical systems perspective on agent-environment interaction'. *Artificial Intelligence* 72, 173-215.

Berk, L. & Garvin, R. (1984). 'Development of Private Speech among Low-Income Appalachian Children'. *Developmental Psychology 20*(2): 271-286.

BonJour, L. (1980). 'Externalist Theories of Empirical Knowledge'. *Midwest Studies in Philosophy*. V.

Bressler, S., and Kelso, J. (2001). 'Cortical coordination dynamics and cognition'. *Trends in Cognitive Sciences* 5, 26–36.

Bressler, S. L. (2002). 'Understanding cognition through large-scale cortical networks'. *Current Directions in Psychological Science,* 11, 58–61.

Brewer, F. W. & Lambert, B. L. (2001). 'The Theory Ladeness of Observation and the Theory-Ladeness of the Rest of the Scientific Process'. *Philosophy of Science,* Vol. 68, No. 3, Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers, pp. S176-S186.

Burge, T. (1993). 'Content Preservation', *The Philosophical Review*, 102, 457-88.

—— (1986). 'Individualism and psychology', *Philosophical Review,* 95: 3-45.

Busemeyer, J., and Townsend, J. (1995). 'Decision field theory'. In *Mind as Motion*, R. Port and T. van Gelder (eds). Cambridge, Mass.: MIT Press.

Busemeyer, J., Townsend, J., T., and Stout, J. (2002). 'Motivational underpinnings of utility in decision making'. In *Emotional Cognition*, S. Moore and M. Oaksford (eds). Philadelphia: John Benjamins.

Butler, K. (1998). *Internal Affairs: A Critique of Externalism In The Philosophy of Mind.* Dordrecht, The Netherlands: Kluwer.

Chisholm, R., M. (1977). *The Theory of Knowledge.* Englewood Cliffs, NJ: Prentice Hall.

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of the Mind.* Cambridge: Cambridge University Press.

—— (1988). 'Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor', *Philosophy of Science* 55: 167-187.

—— (1989). *A Neurocomputational Perspective. The Nature of Mind and the Structure of Science.* Cambridge: The MIT Press.

Clark, A., & Chalmers, D. (1998). 'The Extended Mind'. *Ananlysis* 58, no. 1: 7-19.

Clark, A., (1998). 'Magic Words, How Language Augments Human Computation'. In *Language and Thought: Interdisciplinary Themes.* (1998). P. Carruthers and J. Boucher (Eds). Cambridge University Press: Cambridge.

—— (2001). 'Reasons, Robots, and the Extended Mind', *Mind and Language,* 16;2: 121-145.

—— (2006). 'Soft selves and Ecological Control'. In *Distributed Cognition and the Will.* D. Spurrett, D. Ross, H. Kincaid and L. Stephens (eds). MIT Press, Camb. MA

—— (2007). 'Curing Cognitive Hiccups: A Defense of the Extended Mind', *The Journal of Philosophy,* 104: 163-192. Also available at http://hdl.handle.net/1842/1719 .

—— (2008). *Supersizing The Mind.* Oxford University Press.

—— (2010*a*). 'Memento's Revenge: The Extended Mind, Extended'. In the *Extended Mind.* (2010), Menary (ed.) Cambridge, Massachusetts, MIT press.

—— (2010*b*). 'Coupling, Constitution, and the Cognitive Kind: A Reply to Adams and Aizawa'. In *The Extended Mind.* (2010), Menary (ed.) Cambridge, Massachusetts, MIT press.

Crutchfield, J. (1998). 'Dynamical embodiments of computation in cognitive processes'. *The Behavior and Brain Sciences*, 21, 635.

Dale, R., and Spivey, M. (2006). 'Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation'. *Language Learning*, 56, 391–430.

Di Paolo, E., A. (2003). 'Organismically-inspired robotics: Homeostatic adaptation and natural teleology beyond the closed sensorimotor loop'. In *Dynamical Systems Approach to Embodiment and Sociality*, K. Murase and T. Asakura (eds). Adelaide: Advanced Knowledge International.

Dretske, F. (1970), 'Epistemic Operators', *Journal of Philosophy,* 67, 1007-23.

Davidson, D. (1970). 'Mental Events'. Pp. 79-101 in *Experience and Theory.* Lawrence Foster and J. W. Swanson (eds). Amherst: University of Massachusetts Press.

Estany, A. (2001). 'The Thesis of Theory-Laden Observation in the Light of Cognitive Psychology'. *Philosophy of Science,* Vol. 68, No.2 (Jun ., 2001), pp. 203-217.

Faulkner, P. (2000). 'The Social Character of Testimonial Knowledge'. *The Journal of Philosophy.* 97, 581-601.

Feyerabend, P. K. (1975). *Against Method: Outline of an anarchistic theory of knowledge.*

Fodor, J. (1984). 'Observation Reconsidered', Philosophy of Science, 51: 23-43.

—— (1988). 'A Reply to Churchland's 'Perceptual Plasticity and Theoretical Neutrality'', *Philosophy of Science* 55: 188-198.

—— (1974). 'Special Sciences (or: The disunity of Science as a Working Hypothesis)'. *Synthese* 28: 97-115.

Fricker, E. (1994). 'Against Gullibility'. In *Knowing from Words.* B. K. Matilal & A. Chakrabarti (eds.) 125-61. Dodrecht: Kluwer.

Gaines, B., R. (1994). 'The Collective Stance in Modeling Expertise in Individuals and Organizations'. *Int. J. Expert Systems* 71, 22-51.

Gallagher, S. (2005). *How the Body Shapes the Mind.* Oxford: Oxford University Press.

Gettier, E. (1963). 'Is Justified True Belief Knowledge?', *Analysis* 23, pp. 121-3.

Giere, R. (2002*a*). 'Discussion Note: Distibuted Cognition in Epistemic Cultures'. *Philosophy of Science*, 69.

—— (2002*b*). 'Scientific Cognition as Distributed Cognition'. In *Cognitive Bases of Science*, eds. Peter Carruthers, Stephen Stitch and Michael Siegal, Cambridge: Cambridge University Press, 2002.

—— (2006). 'The Role of Agency in Distributed Cognitive Systems'. *Philosophy of Science,* 73, pp. 710-719.

—— (2007). 'Distributed Cognition without Distributed Knowing'. *Social Epistemology*. Vol. 21, No. 3, pp. 313-320.

Giere, R. & Moffat, B. (2003). 'Distributed Cognition: Where the Cognitive and the Social Merge'. *Social Studies of Science*. 33/2, pp. 1-10.

Goldberg, C., S. (2010). *Relying on Others*. Oxford University Press.

Goldin-Meadow, S. (2003). *Hearing Gesture: How Our Hands Help Us Think.* Harvard University Press, Cambridge. MA, 2003.

Goldman, A. (1976). 'Discrimination and Perceptual Knowledge', *Journal of Philosophy* 73, pp. 771-91.

—— (2009). 'Why Social Epistemology Is Real Epistemology'. In *Social Epistemology,* A. Haddock, A. Millar & D. H. Pritchard (eds). Oxford University Press

Greco, J. (1999). 'Agent Reliabilism', in *Philosophical Perspectives 13:* Epistemology (1999). James Tomberlin (ed.), Atascadero, CA: Ridgeview Press, pp. 273-296.

—— (2004). 'Knowledge As Credit For True Belief', in *Intellectual Virtue: Perspectives from Ethics and Epistemology*. M. DePaul & L. Zagzebski (eds.), Oxford: Oxford University Press.

—— (2007) 'The Nature of Ability and the Purpose of Knowledge', *Philosophical Issues* 17, pp. 57- 69.

—— (2008). 'What's Wrong with Contextualism?, *The Philosophical Quarterly* 58, pp. 299-302.

—— (2010). *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity.* Cambridge University Press.

Hanson, N. R. (1961). *Patterns of Discovery*. Cambridge: Cambridge University Press.

—— (1969). *Perception and Discovery; An Introduction to Scientific Inquiry.* San Francisco: Freeman, Cooper.

Hardwig, J. (1985). 'Epistemic Dependence'. *The Journal of Philosophy*, 82: 335-349.

Harvey, I., Husbands, P., and Cliff, D. (1994). 'Seeing the light: Artificial evolution, real vision'. In *From Animals to Animats 3*, D. Cliff, P. Husbands, J.-A. Meyer, and S. W. Wilson (eds). Cambridge, Mass.: MIT Press.

Harvey, I., Husbands, P., Cliff, D., Thompson, A., and Jakobi, N. (1997). 'Evolutionary robotics: The Sussex approach'. *Robotics and Autonomous Systems*, 20, pp. 205–224.

Heil, J. and Mele, A. (1993). *Mental Causation*. Oxford: Clarendon Press.

Hempel, C. (1966). *Philosophy of Natural Science.* Englewood Cliffs, NJ: Prentice Hall.

—— (1970). *Aspects of Scientific Explanation.* New York: The Free Press.

Heylighen, F., Heath, M., Van Overwalle, F. (2007). 'The Emergence of Distributed Cognition: A Conceptual Framework'. In Proceedings of collective intentionality IV (2004), Volume: IV, Publisher: University of Siena.

Hume, D. (1977). *An Enquiry Concerning Human Understanding,* E. Steinberg (ed.), Indianapolis: Hackett.

Hurley, S. (2010). 'The Varieties of Externalism'. In the *Extended Mind.* (2010), Menary (ed.) Cambridge, Massachusetts, MIT press.

—— (1998). *Consciousness in Action*. Cambridge, MA: Harvard University Press.

Husbands, P., Harvey, I., and Cliff, D. (1995). 'Circle in the round: State space attractor for evolved sight robots'. *Journal of Robotics and Autonomous Systems*, 15, pp. 83–106.

Hutchins, E. (1995). *Cognition in the Wild.* Cambridge: MIT Press.

Kallestrup, J. (2006). 'The Causal Exclusion Argument'. *Philosophical Studies,* Volume 131, No. 2, 459-485.

Kallestrup J. & Pritchard D. H. (*forthcoming*), 'Virtue Epistemology and Epistemic Twin Earth', *The European Journal of Philosophy.*

Kelso, J. A. S., and Engstrøm, D. (2006). *The Complementary Nature.* Cambridge, Mass.: MIT Press.

Kim, J. 1993. *Supervenience and Mind*. New York: Cambridge University Press.

—— 1989. 'Mechanism, Purpose, and Explanatory Exclusion'. *Philosophical Perspectives, 3*.

Knorr-Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge.* Harvard University Press.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* Chicago: The University of Chicago Press.

Lackey, J. (2008). *Learning From Words: Testimony as a Source of Knowledge,* Oxford: Oxford University Press.

—— (2007). 'Why We Do not Deserve Credit for Everything We Know', Synthese 158, pp. 345-61.

Lakatos, I. (1970). 'Falsification and the Methodology of Scientific Research Programmes'. In *Criticism and the Growth of Knowledge.* Imre Lakatos Alan Musgrave (eds.). Cambridge University Press, 1970.

Latour, B. (1999). *Pandora's Hope: Essays on the Reality of Science Studies.* Cambridge, MA: Harvard University Press.

—— (1986). 'Visualization and Cognition: Thinking with Eyes and Hands'. *Knowledge and Society* 6: 1-40.

Lee, D., N. (2006). 'How movement is guided'. Retrieved June 2012 from http://www.pmarc.ed.ac.uk/ideas/pdf/HowMovtGuided100311.pdf

List, C. & Pettit, P. 2010. 'Group Agency and Supervenience'. *The Southern Journal of Philosophy*, Vol. 44, Issue S1, Spring 2010, 86-105.

Logan, K. R. (2003). 'The Extended Mind: Understanding Language and Thought in Terms of Complexity and Chaos Theory'. In *Humanity and the Cosmos*, Daniel McArthur & Cory Mulvihil (eds). Also available at http://www.upscale.utoronto.ca/PVB/Logan/Extended/Extended.html

—— (2006). 'The Extended Mind Model of the Origin of Language and Culture', in *Evolutionary Epistemology and Culture,* N. Gontier et al. (eds), Printed in Netherlands. 149-167.

—— (2008). *The Extended Mind: The Emergence of Language, the Human Mind and Culture.* University of Toronto Press.

Martens, B. (2004). 'The Cognitive Mechanics of Economic Development and Social Change' (PhD thesis, Vrije Universiteit Brussel).

Marsh, K. L., Richardson, M., J., Baron, R., M., and Schmidt, R., C. (2006). 'Contrasting approaches to perceiving and acting with others'. *Ecological Psychology*, 18, pp. 1–37.

McClelland et al. (1986). McClelland, J.L., Rumelhart, D., E., and the PDP Research Group (eds.), *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, vol. 2 .Cambridge, MA: MIT Press.

McKinstry, C., Dale, R., and Spivey, M. (2008). 'Action dynamics reveal parallel competition in decision making'. *Psychological Science*, 19, 22–24.

Menary, R. (2006). 'Attacking the Bounds of Cognition', *Philosophical Psychology.* Vol. 19, No. 3, June 2006, pp. 329-344.

—— (2007). *Cognitive Integration: Mind and Cognition Unbound.* Palgrave McMillan.

Noë, A. (2004). *Action in Perception.* Cambridge, MA:MIT Press.

Nozick, R. (1981), *Philosophical Explanations*. Oxford University Press, Oxford.

Oullier, O., de Guzman, G., C., Jantzen, K., J., Lagarde, J., F., and Kelso, J., A. (2005). 'Spontaneous interpersonal synchronization'. In *European Workshop on Movement Sciences*: Mechanics-Physiology-Psychology, C. Peham, W. Scho¨llhorn, and W. Verwey (eds). Cologne: Sportverlag.

Palermos, S., O. (2010), 'Dualism in the Epistemology of Testimony and the Ability Intuition'. *Philosophia,* Vol. 39, No. 3, pp. 597-613.

—— (2011). 'Belief-Forming Processes, Extended', *Review of Philosophy and Psychology*, Vol. 2, No. 4, pp. 741-765.

—— (*forthcoming*). 'Could Reliability Naturally Imply Safety?', *European Journal of Philosophy.*

Petitot, J. (1995). 'Morphodynamics and Attractor Syntax'. In *Mind as Motion: Explorations in the Dynamics of Cognition.* Port, R., F. and van Gelder, T. (eds). MIT press.

Plantinga, A. (1993). *Warrant and Proper Function.* New York: Oxford University Press.

Pollack, J. (1990), 'Recursive Distributed Representations', *Artificial Intelligence.* 77-105.

Port, R. (2003). 'Meter and speech'. *Journal of Phonetics*, 31, 599–611.

Pritchard, D. H. (*forthcoming*). 'Anti-Luck Virtue Epistemology', *Journal of Philosophy.*

—— (2010*a*). 'Knowledge and Understanding', in A. Haddock, A. Millar & D. H. Pritchard, *The Nature and Value of Knowledge: Three Investigations*, Oxford: Oxford University Press.

—— (2010*b*). 'Cognitive Ability and the Extended Cognition Thesis'. *Synthese.*

—— (2009). *Knowledge,* London: Palgrave Macmillan.

—— (2008), 'Sensitivity, Safety, and Anti-Luck Epistemology'. In *The Oxford Handbook of Skepticism,* J. Greco (ed.), (Oxford: Oxford University Press).

—— (2002), 'Resurrecting the Moorean Response to Scepticism', *International Journal of Philosophical Studies* 10, 283-307.

Putnam, H. (1975). 'The Meaning of "Meaning"'. In *Language, Mind and Knowledge.* K. Gunderson (ed.). Minneapolis: University of Minnesota Press.

Reid, T. (1983). *Essay on the Intellectual Powers of Man*, in R. E. Beanblossom & K. Lehrer (eds.), *Thomas Reid's Inquiry and Essays,* Indianapolis: Hackett.

Richardson, M., J., Marsh, K., L., and Schmidt, R., C. (2005). 'Effects of visual and verbal couplings on unintentional interpersonal coordination'. *Journal of Experimental Psychology: Human Performance and Perception,* 31, pp. 62–79.

Richardson, D., Dale, R., and Kirkham, N. (2007). 'The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue'. *Psychological Science*, 18, pp. 407–413.

Richardson, M., Marsh, K., Isenhower, R., Goodman, J., and Schmidt, R. (2007). 'Rocking together: Dynamics of intentional and unintentional interpersonal coordination'. *Human Movement Science*, 26, pp. 867–891.

Roe, R. M., J. Busemeyer, and J. T. Townsend (2001). 'Multialternative decision field theory: A dynamic artificial neural network model of decision making'. *Psychological Review*, 108, pp. 370–392.

Ross, D. and Ladyman, J. (2010). 'The Alleged Coupling-Constitution Fallacy and the Mature Sciences'. In *The Extended Mind.* (2010), Menary (ed.) Cambridge, Massachusetts, MIT press.

Rowlands, M. (1999). *The Body in Mind: Understanding Cognitive Processes.* New York: Cambridge University Press.

—— (2009). 'Extended Cognition and the Mark of the Cognitive'. *Philosophical Psychology,* 22(1); pp. 1-19.

Rupert, D. R. (2004). 'Challenges to the Hypothesis of Extended Cognition'. *Journal of Philosophy,* 101: 389-428.

—— (2009). *Cognitive Systems and the Extended Mind.* Oxford University Press.

—— (2010). 'Representation in Extended Cognitive Systems: Does the Scaffolding of Language Extend the Mind?'. In the *Extended Mind.* (2010), Menary (ed.) Cambridge, Massachusetts, MIT press.

Sawyer, K. (2001). 'Emergence in Sociology: Contemporary Philsophy of Mind and Some Implications for Sociological Theory'. *The American Journal of Sociology*. Vol. 107, No.3, pp. 551-585.

Schmidt, R., Carello, C., and Turvey, M. (1990). 'Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people'. *Journal of Experimental Psychology: Human Perception and Performance,* 16, pp. 227–247.

Simon, H., A. (1969). 'The Architecture of Complexity'. Pp. 192-229 in *The Sciences of the Artificial,* by Herbert A. Simon. Cambridge, Mass.: MIT Press.

Sosa, E. (1988).'Beyond Skepticism, to the Best of our Knowledge'. *Mind,* New Series, vol. 97, No.386, pp. 153-188

—— (1993). 'Proper Functionalism and Virtue Epistemology'. *Nous,* Vol. 27, No. 1, 51-65.

—— (1999). 'How to Defeat Opposition to Moore', *Philosophical Perspectives,* 13, pp. 141-54.

—— (2000). 'Skepticism and Contextualism', *Philosophical Issues* 10, pp. 1-18.

—— (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge,* Oxford: Clarendon Press.

Spivey, M., and Dale, R. (2006). 'Continuous temporal dynamics in real-time cognition'. *Current Directions in Psychological Science*, 15, 207–211.

Spivey, M. (2007). *The Continuity of the Mind*. Oxford University Press.

Sprevak, M. (2010). 'Inference to the hypothesis of extended cognition'. *Studies in History and Philosophy of Science,* 41: 353-362.

Stephen, D., D., J., and Isenhower, R. (2007). 'Dynamics in development: New structure through self-organization'. Paper presented at the 14th International Conference on Perception and Action.

Steup, M. (1999). 'A Defense of Internalism'. In L. Pojman. Editor. *The Theory of Knowledge: Classical and Contemporary Readings*. 2nd edition. Belmont. Wadsworth Publishing.

Sutton, J., Barnier, A., Harris, C., Wilson, R. (2008). 'A conceptual and empirical framework for the social distribution of cognition: The case of memory'. *Cognitive Systems Research,* Issues 1-2, pp. 33–51.

Teller, P. 1992. 'A Contemporary Look at Emergence'. Pp. 139-53, in *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism,* A. Bechermann, H. Flohr, and J. Rim. Berlin (eds): Walter de Gruyter.

Theiner, G. (*manuscript*). Cognitive Processes, Cognitive Systems, and the Extended Mind.

Theiner, G. & Allen, C. & Goldstone, R. (2010). 'Recognizing Group Cognition'. *Cognitive Systems Research*, Vol. 11, Issue 4, pp. 378-395.

Thompson, E., and Varela, F. (2001). 'Radical Embodiment: Neural dynamics and consciousness'. *Trends in Cognitive Sciences*, 5, pp. 418–425.

Tollefsen, D. 2002. 'Organizations as True Believers', *Journal of Social Philosophy,* Vol. 33 No. 3, Fall 2002, pp. 395-410.

Turvey, M. T., and Moreno, M. (2006). 'Physical metaphors for the mental lexicon'. *Mental Lexicon*, 1, pp. 7–33.

Van Orden, G., Holden, J., and Turvey, M. (2005). 'Human cognition and 1/f scaling'. *Journal of Experimental Psychology: General*, 134, pp. 117–123.

Van Gelder, T. (1995). 'What Might Cognition Be, If Not Computation?'. *The Journal of Philosophy,* Vol. 92, No. 7 (Jul., 1995), pp. 345-381.

Varela, F., Lachaux, J., P., Rodriguez, E., and Martiniere, J. (2001). 'The brainweb: Phase synchronization and large-scale integration'. *Nature Reviews Neuroscience*, 4, pp. 229–239.

Varela, F. (1979). *Principles of Biological Autonomy.* New York. North-Holland.

Wegner, M., Giuliano, T., Hertel, P. (1985). 'Cognitive interdependence in close relationships'. In W. J. Ickes (Ed.), *Compatible and incompatible relationships* (pp. 253–276). New York: Springer-Verlag.

Weiner, M. (2003). 'Accepting Testimony'. *The Philosophical Quarterly* 53, 256-64.

Wheeler, M. (2004). 'Is Language the Ultimate Artifact?', *Language Sciences* 26: 693-715.

—— (2005). *Reconstructing the Cognitive World*. MIT Press, Cambridge, Massachusetts.

Wilson, R. A. (2000). 'The mind beyond itself'. In D. Sperber (ed.), *Metarepresentations: A Multidisciplinary Perspective,* New York University Press, pp. 31-52.

—— (2004). *Boundaries of the Mind: The individual in the Fragile Sciences: Cognition.* New York: Cambridge University Press.

Wilson, R., A. (2005). 'Collective Memory, Group Minds, and the Extended Mind Thesis'. *Cognitive Processing*, Vol. 6, Issue 4, pp. 227-236.

—— (*in press*). 'Extended Vision'. In *Perception, action, and consciousness*. N. Gangopadhyay, M. Madary & F. Spicer (eds). Oxford: Oxford University Press.

Wimsatt, W. S. (1986). 'Forms of Aggregativity'. In *Human Nature and Natural Knowledge.* M. G. Grene, A. Donagan, A. N. Perovich, & M. V. Wedin (Eds), (pp. 259-291). Dordrecht: Reidel.