Social machines

(*Forthcoming* in *Phenomenology and the Cognitive Sciences*)

SOCIAL MACHINES: A PHILOSOPHICAL ENGINEERING

S. Orestis Palermos

University *of* Edinburgh

Abstract. In *Weaving the Web* (2000), Berners-Lee defines Social Machines as biotechnologically hybrid Web-processes on the basis of which, "high-level activities, which have occurred just within one human's brain, will occur among even larger more interconnected groups of people acting as if the shared a larger intuitive brain" (201-202). The analysis and design of Social Machines has already started attracting considerable attention both within the industry and academia. Web science, however, is still missing a clear definition of what a Social Machine is, which has in turn resulted in several calls for a "philosophical engineering" (Halpin 2013; Hendler & Berners-Lee, 2010; Halpin et al., 2010). This paper is a first attempt to respond to this call, by combining contemporary philosophy of mind and cognitive science with epistemology. The idea of *philosophical engineering* implies that a sufficiently good conception of Social Machines should be of both theoretical and practical advantage. To demonstrate how the present approach can satisfy both objectives it will be used in order to address one of Wikipedia's (the most famous Social Machine to date) most worrying concerns—i.e., the current and ongoing decline in the number of its active contributors (Halfacker et al., 2012).

1. Introduction

Web science is a fast emerging field (Berners-Lee et al., 2006), which, just as any other nascent discipline, is permeated by philosophical discourse (Kuhn 1962). On one hand, computer scientists have made several attempts to explore the relevance of contemporary philosophy of mind and cognitive science to the understanding of the nature and future of the Web (Halpin et al., 2010; Halpin, 2013; Smart, 2012; Smart and Shadbolt, 2014; Smart, 2014; Smart et al. forthcoming). On the other hand, philosophers have already started discussing some of the challenges that the expansion of the Web may pose. Typically, at the forefront of such discussions are "issues concerning knowledge, identity and trust, […] the understanding of authorship within new collaborative and collective creative ensembles, and the relation between, on the one hand, the structure and functioning of the Web and on the other hand, the strengths and weaknesses of basic biological cognition (Halpin, Clark and Wheeler, 2010).

Before addressing such questions, however, Web science first needs to get clear about an important, perhaps fundamental, issue concerning the past, present and future of the Web—namely, the flow of information and its direction. In *Weaving the Web* (Berners-Lee et al. 2000)—published some 15 years ago, when the Web had been around for hardly a decade's time—Berners-Lee accentuated the fact that the Web (at least back then) was still a work in progress...that it still had quite some way to go before achieving its final goal of being an "intimate collaborative medium" (57). The problem, noted Berners-Lee, was that "as soon as developers got their client working as a browser and released it to the world, very few bothered to continue to develop it as an editor" (57). And this run contrary to the vision of "the information space as something to which everyone has immediate and intuitive access, and not just to browse, but to create" (157).[1]

Looking at today's Web, it is apparent that Berners-Lee's vision has had a strong impact on it. At its 25th anniversary, the Web has grown into what is known as *Web 2.0*, characterized by "greater levels of user participation in the creation, maintenance and editing of online "content" (Smart 2014). In return, increased user participation has led individuals to share information and interact in new and previously unimagined ways, giving rise to a distinguishing feature of Web 2.0, namely the *'Social Web'*: A suite of applications, services, technologies, formats, protocols and other resources, all united in their attempt to both foster and support social interaction" (Smart, 2014).

The Social Web is a transparent and rather mundane aspect of everyday digital life, with most of us participating in it without even realizing it. Nevertheless, it is also a fast evolving process with the potential to fundamentally transform human society. As Berners-Lee and Hendler note in the opening of an article that was written 10 years after the publication of *Weaving the Web,* "much has been written about the profound impact that the World Wide Web has had on society. Yet, it is primarily in the past few years, as more interactive 'read/write' technologies (e.g., Wikis, blogs and photo/video sharing) and social networking sites have proliferated, that the truly profound nature of this change is being felt" (Hendler & Berners-Lee, 2010).

But has the Web unleashed its full potential yet? To answer this question, it is instructive to consider a famous passage from Berners-Lee's book, where the idea of 'Social Machines' is introduced for the first time. Even though he does not provide much detail about this emerging concept, Berners-Lee makes three definitional remarks of increasing importance:

[1] Remarks to the same effect can be found in several passages throughout the book. See for example pp. 63 and 169.

(i) Social Machines are Web-driven processes in which "the people do the creative work and the machine does the administration" (172).

(ii) Their distinctive value is that they "will enable us to do things we just couldn't do before" (174).

(iii) This will become possible because, on the basis of Social Machines, "creativity will arise across larger and more diverse groups, [and] high-level activities, which have occurred just within one human's brain, will occur among even larger more interconnected groups of people acting as if they shared a larger intuitive brain" (201-202).

These remarks are of increasing importance, because, even though (i) marks an *organisational* difference between Social Machines and existing social systems and (ii) puts forward a promise about their importance, in (iii), Berners-Lee introduces a bold analogy in order to explain the functional origins of the distinctive value of Social Machines: Social Machines are expected to provide unprecedented opportunities, because they will allow groups of people to operate as if they were parts of, or gave rise to, a (super) brain.

Two specific questions immediately spring to mind: First, are we there yet? And secondly, how can one achieve such a goal—how can we efficiently design Social Machines? Research, spread over several institutions around the world, has already started turning its attention to such and similar questions.[2] Nevertheless, as Smart and Shadbolt (2014) note, this research is paradoxically missing a clear definition of what Social Machines are supposed to be. For the time being, Social Machines are picked out by *ostension* and a list of potential candidates includes Facebook, mySpace, Twitter, Galaxy Zoo, Wikipedia, Intellipedia and so on.[3] But it is an open question whether these examples are truly representative of Social Machines. Identification by ostension, after all, can be particularly unreliable—especially when the relevant term is entirely new. In fact, if the aim is to distinguish Social Machines from other forms of social computing, then, given Berners Lee's initial definition—which

[2] Take for example SOCIAM (http://sociam.org)--a collaboration between Southampton, Oxford and Edinburgh universities, set to explore the nature of Social Machines. Or consider the Smart Society Project (http://www.smart-society-project.eu)--a collaboaration between the University of Trento, University of Edinburgh, U-Hopper s.r.l., German Research Centre for Artificial Intelligence, University of Oxford, Ben-Gurion University of the Negev, Imaginary s.r.l, University of Karlstad, Vienna University of Technology and University of Southampton, whose goal is to understand and design hybrid systems where people and machines can work tightly together to build smarter societies.

[3] Facebook, mySpace and twitter are social networking websites where users can post, share and 'like' comments and online content. Galaxy Zoo is a citizen science project that enlists members of the public to assist with the morphological classification of large number of galaxies. Wikipedia is a free online encyclopedia that allows its users to edit almost any article accessible. Intellipedia is an online system for collaborative data sharing used by the United States Intelligence Community.

appeals to the functional organization of a human brain—many of the above examples may not be profitably seen as Social Machines.

Pondering on the nature of Social Machines, Smart and Shadbolt (2014) have recently attempted to improve on Berners Lee's initial remarks by breaking them down in the following three components: Social Machines refer to a process which (1) multiple individuals engage in; (2) it is "biotechnologically hybrid in nature (i.e., it requires the contribution of both machine and human elements)"; (3) it involves a "commitment to the idea of human and technological elements fulfilling particular kinds of roles, roles that are (perhaps broadly) construed as either creative or administrative in nature."

This is a helpful reformulation of Berners-Lee's idea, but it does not exclude any of the above examples from counting as Social Machines. The reason is that it ascribes no weight on Berners-Lee's point regarding the emergence of a larger brain consisting of all the underlying components. Given how Berners-Lee intended the term, this seems to be a distinctive feature of Social Machines.

Smart and Shadbolt recognize this when they suggest that the components of Social Machines should operate as parts of a single unified system. Accordingly, in a final attempt to offer a definition they note:

> Social Machines can be defined as web-based socio technical systems in which the human and technological elements play the role of participant machinery with respect to the mechanistic realization of system-level processes. (2014).

This suggestion is in line with Berners-Lee's initial idea, as it does stress that Social Machines should function as overarching unified systems. The problem, however, is that, taken on its own, it is not a particularly clear definition, due to a lack of understanding with regards to the underlying concepts of 'participant machinery' and 'system-level processes'. In an attempt to define these concepts, Smart and Shadbolt cite Clark (2008, 208), but Clark does not go into the details of what may count as 'participant machinery' and 'system level processes' either. All we are provided with, instead, is the following descriptive but not explanatory note: 'Participant machinery' means "to form part of the very machinery by means of which mind and cognition are physically realized and hence to form part of the local material supervenience base for various mental states and processes." (Clark 2008, 208).

Explaining therefore what may count as a Social Machine in Berners-Lee's intended use of the term requires an explanation of what may count as part of the machinery that gives rise to mental states and cognitive processes. To answer this central question, this paper focuses on the concept of *cognitive integration* as it has been introduced within philosophy of mind and epistemology. The reason for this move is straightforward: Within both fields, the concept of *cognitive integration* has been introduced in order to account for precisely the same

question that an adequate definition of Social Machines has so far been missing—i.e., what may count as proper part of a cognitive system?

As we shall see, answering this question has given rise to the possibility of cognitive systems that may be distributed between several individuals and their artifacts at the same time—an idea that is very close, if not identical, to the way Berners-Lee has envisioned Social Machines. Focusing on the concept of *cognitive integration* can therefore help understand what is required from a component in order to count as 'participant machinery' of an overall cognitive system. Given Smart and Shadbolt's working definition, this is key to building socio-technical systems that can function as distributed, unified brains.

Of course, one may argue that this is a bold analogy that Web science needs not take seriously. Perhaps the term Social Machine should (continue to) be loosely used to refer merely to whatever falls under the broad category of 'social computing' or whatever embodies the three properties that Smart and Shadbolt point out above. But what if philosophy of mind and epistemology can offer a promising way to make sense of Berners-Lee's aspiration of 'Social Machines' as 'interconnected groups of people acting as if they shared a larger intuitive brain'? What if such an approach could make a substantial difference to the efficient design of Social Machines? Recent calls for a "philosophical engineering" (Halpin 2013; Hendler & Berners-Lee, 2010; Halpin et al. 2010) seem to suggest that this is an open possibility indeed.

The emerging idea of 'philosophical engineering' fits well with contemporary, naturalist philosophy (Papineau 2015). Just as philosophical naturalism brings philosophy and the natural sciences closer by facilitating and motivating a constant feedback between the two, similarly, the idea behind philosophical engineering is to motivate the exchange of ideas between philosophy and engineering. Analogies, concepts, even models that originate from a number of engineering disciplines have so far had a resounding impact on debates within philosophy of mind, philosophy of biology and philosophy of science. Why couldn't the effect run the other way too?

'Philosophical engineering' is the idea that philosophy has the potential to impact future engineering choices regarding the efficient design of emerging technologies. Coming up with capacity enhancing artifacts and infrastructures presupposes a sophisticated understanding of the user's mind as well as the nature of society. Philosophy can be particularly revealing on both fronts.

Halpin et al. (2010) clearly acknowledge this potential when it comes to designing Web technologies:

> From Wikipedia to Google, [the] ability to trust information on the Web is one of the most pressing problems facing the Web today [(O'Hara & Hall, 2008)], so a detailed philosophical

> analysis of this topic promises to have a significant impact on the practice of Web engineers designing these systems

Similarly, with respect to building Social Machines, Hendler and Berners-Lee note (2010, p. 5) that "we must bring to bear not only the best engineering and theoretical perspectives of current computer science, but also to create new and exciting theory and technology as we forge the path towards our goal." We need to find a way to

> "create and then evolve new kinds of social machines that will provide people, individually and collectively, with the ability to immerse themselves in the accumulated knowledge and the constant interactions of humankind. People's interactions will be not just as passive recipients of information created by others, but also as contributors to this global information space in a way far beyond that of today's Web" (5).

This paper is a first attempt to respond to this call for *philosophically informed engineering*. By combining contemporary philosophy of mind and cognitive science with epistemology, it lays out a framework for defining and engineering current and future Social Machines, with a special focus on the concept of *cognitive integration*. Moreover, the idea of *philosophical engineering* implies that a sufficiently good conception of Social Machines should be not only of theoretical but of practical advantage too. In order to demonstrate how to put the present approach in practice, towards the end of the paper, I address one of Wikipedia's—the most famous Social Machine to date—most worrying concerns (i.e., the current and ongoing decline in the number of its active contributors). The general aim is to provide at least some indication of how tomorrow's engineers should proceed with the efficient design of Social Machines.

2. Cognitive Integration

The main aim of this paper is to explore the idea that the distinctive feature of Social Machines is their potential to allow high-level cognitive activities to "occur among [...] groups of people acting *as if they shared a larger intuitive brain*" (Berners-Lee et al. 2000, p. 202, emphasis added). This is a bold statement that may be disregarded as a mere metaphor with no practical implications. Nevertheless, recent developments within epistemology and philosophy of mind and cognitive science suggest otherwise. By focusing on the concept of *cognitive integration* a number of authors have suggested that, provided the right interactivity is in place, several individuals and their artifacts can give rise to an overall distributed cognitive system that comprises of all of them at the same time. Such advances demonstrate that not only is it possible to take Berners-Lee's remarks seriously, but that they can also be used to provide specific guidance to the design of future Social Machines. To demonstrate how this is possible, the following subsections focus on the central concept of *cognitive integration* both from the perspective of philosophy of mind and epistemology.

2.1 Cognitive Integration in Philosophy of Mind and Cognitive Science

Berners-Lee's vision to generate a group of people that acts as a unified brain as well as his remarks that in such cases "people's interactions [should] be not just as passive recipients of information created by others, but also as contributors" (Hendler and Berners-Lee, 2010, p. 5) bring to mind what philosophers of mind and cognitive scientists refer to as *active externalism*.

Active externalism is an emerging approach to cognition, which centers around the following claim: If natural, social or technological aspects of the environment are heavily involved in *driving* some cognitive process as it unfolds over time, then these environmental factors should be thought of as proper, *constitutive* parts of the relevant cognitive system. Active externalism has appeared in the literature under several labels and formulations—e.g., the extended mind thesis (Clark and Chalmers 1998), cognitive integration (Menary 2007), environmentalism (Rowlands 1999), location externalism (Wilson 2000, 2004), the hypothesis of extended cognition (Clark and Chalmers 1998), the hypothesis of distributed cognition (Hutchins 1995, Sutton, 2008; Barnier et. al, 2008; Theiner et al., 2010; Theiner and O-Connor, 2010, etc.). For present purposes, we can concentrate on the latter two formulations alone.

Focusing on cognitive processing, the hypothesis of extended cognition is the claim that "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. 2). Similarly, the hypothesis of distributed cognition (Barnier et al., 2008; Heylighen et al., 2007; Hutchins, 1995; Sutton et al., 2008; Sutton, 2008; Theiner et al. 2010; Theiner and Goldstone, 2010; Tollefsen & Dale, 2011; Tollefsen, 2006; Wilson, 2005) holds that cognitive processing may not just be extended beyond the agent's head or organism but might be even distributed amongst several individuals along with their epistemic artifacts. Accordingly, though strictly speaking distinct, the hypothesis of distributed cognition differs from the hypothesis of extended cognition only in that, in the former case, cognitive processes and the resultant cognitive systems extend to include not only artifacts but other individuals as well.

With respect to argumentative lines, active externalism—especially in the form of the extended mind thesis—has been traditionally associated with common-sense functionalism (Braddon-Mitchell & Jackson, 2006). It has been recently argued (Chemero 2009, Palermos 2014*a*, Palermos forthcoming), however, that contrary to the extended mind thesis, the focus of the extended and distributed cognition hypotheses is not on mental states (such as beliefs and desires, understood in common-sense functionalist terms), but on extended (and distributed) *dynamical* cognitive processes and the overall cognitive *systems* these processes give rise to. Accordingly, the extended and distributed cognition hypotheses do not need to rely for

their support on common-sense functionalism. Instead, they can be motivated on the basis of Dynamical Systems Theory (DST), which is perhaps the most powerful, if not the only, mathematical framework for studying the behavior of dynamical systems, in general. [4]

According to this conceptual framework, in order to claim that two (or more) systems give rise to some extended or distributed process and, thereby, to an overall extended or distributed system (either way, to a *coupled* system, in DST terms), we need to establish that the contributing parts give rise to a cognitive task by non-linearly interacting with each other on the basis of ongoing feedback loops between them (Chemero 2009, Froese et al. 2013, Sutton et al. 2008, Theiner et al. 2010, Wegner et al. 1985, Tollefsen & Dale 2011, Palermos 2014*a*, Palermos forthcoming). The underlying rational is that non-linear interactions between components give rise to an overall *integrated* system that consists of all them as the same time.

There are two main reasons for postulating this overall *cognitively integrated* system: (1) The aforementioned non-linear interactions give rise to new systemic properties that belong only to the overall system and to none of the contributing subsystems alone (therefore one *has to* postulate the overall extended or distributed system); (2) Said interactions also make it impossible to decompose the two systems in terms of distinct inputs and outputs from the one subsystem to the other (therefore one *cannot but* postulate the overall system).[5] Accordingly, the claim, on the basis of DST, is that in order to have an extended or even distributed cognitive system—as opposed to merely an embedded one (cf. Adams & Aizawa, 2001, 2010; Rupert, 2004, 2009)—the contributing members (i.e., the relevant cognitive agents and their artifacts) need to give rise to a cognitive task by *non-linearly interacting* (on the basis of ongoing feedback loops) with each other.[6]

In practice this means that asking for directions from a stranger on the street or receiving testimony in the court of law cannot give rise to a distributed cognitive system. The reason is that in such cases *there are no non-linear interactions* between the cognitive processes of

[4] See also (Shani, 2013), whose view—*viz.*, moderate active externalism—is similar to what we here call the hypothesis of extended cognition (though note that Shani's arguments do not so heavily rely on DST, and his view is stronger than the hypothesis of extended cognition in that it denies—instead of remaining silent on the matter—the extension of (common-sense functionalist) mental states. For more details on why common-sense functionalism is necessary for the extended mind thesis, but not the extended and distributed cognition hypotheses, see (Palermos 2014*a*, Palermos forthcoming). Though note, again, that the hypotheses of extended and distributed cognition are neither incompatible with common-sense functionalism, nor anti-functionalist on the whole. In so far as a cognitive process is a function, these two hypotheses *are* compatible with functionalism.

[5] To preempt a possible worry, here, the relevant reciprocal interactions need only be continuous during the operation of the relevant coupled cognitive system and the unfolding of any processes related to it. For example, if, as part of her job and during normal working hours, individual *S* participates in distributed cognitive system *X*, *S* does not need to continuously interact with the other members of *X*, when she is at home. However, whenever *X* is in operation, *S* must continuously and reciprocally interact with the rest of the *X*-members. For a detailed explanation of why the existence of non-linear relations that arise out of reciprocal interactions between agents and their artifacts ensures the existence of distributed cognitive systems see (Palermos forthcoming).

[6] An anonymous referee worries that the criterion of mutual interactivity on the basis of feedback loops cannot deal with the 'cognitive bloat' worry. That is, if that's all that is required for extended and distributed cognition to arise then we will end up with a cognitive bloat whereby cognition will seem like leaking in implausibly many directions—i.e., we will have to postulate an implausible number of extended and distributed cognitive systems. To solve this problem the referee suggests reintroducing the common sense functionalist criteria of trust, reliability, availability and past endorsement as introduced by Clark and Chalmers (1998). The following paragraph and fn. 7 demonstrate how the ongoing feedback loops criterion can handle the cognitive bloat worry within the context of the hypothesis of distributed cognition without such additional criteria. In (Palermos 2014*a*) there is also a detailed analysis on how the feedback loops criterion suffices, on its own, to overcome the cognitive bloat worry, by providing a clear distinction between the hypothesis of extended cognition and the hypothesis of embedded cognition.

the involved individuals. Specifically, in cases of testimony, even though there is a minimal amount of interpersonal communication and interaction involved, the cognitive processes of the individual that produces the relevant information are not mutually interrelated with the cognitive processes of the individual that receives the information. Instead, there is only one-way, linear dependence between the individuals under consideration, in that the way the speaker formulates the information she delivers is entirely independent of the recipient's cognitive processes—no continuous feedback loops between the two individuals are responsible for generating the content of the relevant information. On the contrary, as it will become apparent later on (see § 2.3), in the case of Transactive Memory Systems (Wegner et al., 1985; Wegner, 1986; Sutton 2008; Barnier et al. 2008) as well as in the case of several scientific research teams (Knorr-Cetina, 1999, Nersessian, 2006; Giere, 2002*a,* 2002*b,* 2003)) the completion of the relevant cognitive task involves ongoing, dense feedback loops between the participating individual members suggesting that the criterion of ongoing mutual interactions is thereby satisfied. Accordingly, in such cases, we can talk of the presence of an overall distributed cognitive system that consists of all the participating individuals at the same time.[7]

The above offers a way to clearly distinguish between distributed cognition and cognition that is merely socially embedded. Nevertheless, in order to properly appreciate the hypothesis of distributed cognition, it is important to also understand the associated concept of *emergence*. Theiner et al., for example, claim that "groups have the potential to display emergent cognitive properties that no individual member has, or might even be capable of having" (Theiner et al., 2010, p. 381). But how are we to understand the nature of *emergent*, irreducible, collective properties and what is their relation to DST?

Emergent, collective properties refer to *regularities* in the behavior of the group as a whole. Each token instance of any such behavior may still, in principle, be performed by a single individual or at least by a random collection of them. But in order for such behavior to

[7] This may raise a possible concern with other forms of interpersonal communication that potentially involve dense mutual interactivity between the involved parties, such as in the case of dialogue. Could such cases of interpersonal communication qualify as instances of collective cognition? Interestingly, a number of authors (Fusaroli, Gangopadhyay and Tylen, 2013; Fusaroli, Raczaszek-Leonardi and Tylen 2013) have recently suggested that dialogue between two individual can, many times, qualify as a case of distributed cognition. This could potentially lead to a cognitive bloat (Rowlands 2009) whereby any form of communication between two individuals would qualify as giving rise to a distributed cognitive system. As noted above, however, not all instances of interpersonal communication will satisfy the criterion of mutual interactivity. Put simply, not all interpersonal communication is a form of dialogue. In their paper, Fusaroli, Gangopadhyay and Tylen provide essentially the same response:

> Not all conversational arrays automatically come to constitute interpersonal synergies. This requires a certain level of skillful linguistic engagement. As evident from the empirical studies reviewed, synergy effects can be achieved to a lower or higher degree. In other words, interlocutors can become more or less close to an ideal model of dialogical mind according to the dymanics at play. Importantly, we also suggest a possible mechanism for the creation and maintenance of dialogical minds: the co-construction of interactional routines, such as the context-sensitive alignment of expressive behavior (Fusaroli, Gangopadhyay and Tylen, 2013).

be *regular*, such that it can count as the property of some system, the group entity must be in place.

To elaborate, any behavior that could be classed as the manifestation of some system's (cognitive) *properties* (such as the set of processes giving rise to a scientific research team's experiment) cannot count as such if it is merely the product of all the necessary ingredients momentarily coming together in a fleeting way. Instead, the relevant behavior needs to arise out of the *coordinated* and (thereby) self-regulatory activity of some appropriate collection of units that will allow it to be (at least potentially) *regular* behavior. On the basis of this point, we can also draw the connection with the preceding discussion on non-linear interactions and DST: The presence of the *coordinated* (non-linear) interactions between the individual members of the group—which is needed in order for their behavior to count not as random behavior but as the property of some system—also renders, according to the DST arguments for positing coupled systems, the postulation of the corresponding group entity necessary.

The idea of how group entities *emerge* can therefore be summarized in the following way: When individual members *coordinate* on the basis of reciprocal interactions, they adapt mutually to each other by *restricting* their actions so as to *reliably*—that is, regularly—achieve ends that they would only luckily—if ever—bring about were they to act on their own. Via the application of such positive mutual constrains, which result from, and further guide, the members' coordinated activity, new collective properties (i.e., regular behaviors) come about and the collective achieves a stable configuration that is necessary for its survival and further development. This process of "self organization and further evolution of the collective" as Heylighen et al. put it (2004, p. 6), "effectively creates a form of 'social' organization in which agents help each other so as to maximize the collective benefit." Moreover, since such self-organization can only arise on the basis of non-linear interactions between the members of the relevant group, DST holds that, in such cases, it is necessary to also postulate the corresponding group entity.

Overall then, short of postulating the relevant collective entity, it is impossible to account for the individual members' restrained behavior: A behavior that results from the members' coordinated activity and which gives rise to emergent properties in the form of unprecedented regularities in the behavior of the group *as a whole*.[8]

2.2 Cognitive Integration in Epistemology

The above presents a broad picture of how philosophy of mind and cognitive science understand the idea of cognitive integration and its relation to the concept of emergence.

[8] For a detailed defense of group properties and entities on the basis of a naturalized version of emergence, see (Palermos forthcoming).

Cognitive integration, however, has also lately come to occupy a central position within epistemology. To appreciate the epistemic import of cognitive integration, it is useful to consider a long-standing problem with the traditional account of knowledge as justified true belief—specifically, a problem with the justification component.

Most epistemologists hold that justification/epistemic responsibility is some form of ability to provide explicit positive reasons in support of our beliefs or in support of the reliability of our beliefs.[9] This is an intuitive way to think about justification/epistemic responsibility but the problem is that there are several belief-forming processes, such as our visual perception and memory, which even though are supposed to be knowledge-conducive, most epistemic agents have no idea how they work or why they are reliable. Accordingly, when we acquire knowledge on their basis, it seems incorrect to require explicit positive reasons in their support.

In order to solve this long standing problem, several epistemologists have recently suggested giving up the aforementioned strong understanding of justification, according to which one should be able to provide explicit positive reasons in support of one's beliefs, and instead embrace a considerably weaker approach to knowledge and justification. According to the weaker alternative, in order for one's true beliefs to qualify as knowledge, they must simply be the product of a belief-forming process that counts as a cognitive ability.[10] This is known as the *ability intuition on knowledge* and its intuitive appeal comes from the fact that cognitive abilities seem to be the sort of belief-forming processes that can generate knowledge, even if one has no explicit positive reasons to offer in their support.[11] After all, no one needs to explain why their vision or hearing is reliable when they come to acquire knowledge on their basis.

Nevertheless, if this is the way to approach knowledge and justification, there seem to be two central questions that we need to further ask: (1) When does a process count as a cognitive ability and thereby as knowledge conducive, and—depending on how we answer (1)—(2) what is the sense in which one can be justified/epistemically responsible on the basis of one's cognitive abilities, but without requiring to offer any explicit reasons in their support?

It is in answering these two important questions that epistemologists have turned to the concept of *cognitive integration*. Greco has recently proposed that in order for a process to

[9] This is known as the access internalist approach to justification. For classic defenses of the view see (BonJour, 1985; Chisholm, 1977; Steup, 1999).

[10] This is the main tenet of what is known within contemporary epistemology as 'virtue reliabilism.' A historically prior but weaker alternative is 'process reliabilism' according to which in order for a belief-forming process to count as knowledge conducive, it only needs to be reliable, independently of whether it may count as a cognitive ability or not (for an overview of reliabilism see Goldman and Beddor 2015). Given that process reliabilism is not concerned with whether the relevant process may count as a cognitive ability, it also neglects the concept of cognitive integration, which is the present paper's main philosophical focus and key for designing Social Machines that could qualify as distributed cognitive systems (see also fn. 13).

[11] The idea that knowledge must be grounded in cognitive abilities can be traced back to the writings of (Sosa 1988, 1993) and Plantinga (1993*a*, 1993*b*). For more recent approaches to this intuition, see Greco (1999; 2004; 2007; 2010) and Pritchard (2009, 2010a, 2010b, 2010c, 2012).

count as a cognitive ability (and thereby as knowledge-conducive) it must have been *cognitively integrated*, where—just like in the case of extended and distributed cognition—"cognitive integration is a function of cooperation and interaction, or cooperative interaction with other aspects of the cognitive system" (2010, 152). Accordingly, the answer to the first question is that, no matter whether a belief-forming process is entirely internal or partly external to the agent's biological organism, it will still count as a cognitive ability (and thereby as knowledge-conducive) provided that it is *cognitively integrated* on the basis of processes of mutual interactions with other aspects of the cognitive system.

One of the virtues of this approach to knowledge and justification is that it is fairly straightforward: In order for a reliable belief-forming process to count as knowledge-conducive, it must also count as a cognitive ability, and, in order for that to be the case, the relevant belief-forming process must mutually interact with other aspects of the cognitive system. Additionally, a further advantage of the approach is that it can also provide a satisfactory response to the second question we posed above—i.e., what is the specific sense in which one can be justified/epistemically responsible on the basis of one's cognitive abilities, even if one has no explicit reasons to offer in support of their reliability? The key, again, is to focus on the cooperative and interconnected nature of cognitive abilities: If one's belief-forming process interacts cooperatively with other aspects of one's cognitive system then it can be continuously monitored in the background such that *if* there is something wrong with it, *then* the agent will be able to notice this and respond appropriately. Otherwise—if the agent has no negative beliefs about his/her belief-forming process—he/she can be subjectively justified/epistemically responsible in employing the relevant process *by default*, even if he/she has absolutely no positive beliefs as to whether or why it might be reliable.

2.3 Cognitive Integration at the Intersection of Epistemology and Cognitive Science

The above brings to the fore the possibility that knowledge-conducive cognitive abilities can be extended to the artifacts we employ or even be distributed between several individuals at the same time. This is because—even though independently developed and motivated within distinct philosophical disciplines—the two theories that the previous subsections review put forward the same condition in order for a process to count as *cognitively integrated* (and thereby, by the lights of virtue reliabilism, as knowledge-conducive): Just as proponents of active externalism claim that a cognitive system is integrated when its contributing parts engage in ongoing reciprocal interactions (independently of *where* these parts may be located), so Greco claims that cognitive integration of a belief-forming process (be it internal or external to the agent's organism) is a matter of cooperative interactions with other parts of the cognitive

system.[12] Accordingly, there is no principled theoretical bar disallowing the belief-forming processes of extended or even distributed cognitive systems from counting as knowledge-conducive.

Specifically, provided that the relevant system is cognitively integrated on the basis of the mutual interactions of its component parts, it can generate epistemically responsible/justified beliefs, independently of whether it is organism-bound, extended or distributed. The reason is that the ongoing interactivity of its component parts—i.e., its cognitively integrated nature—allows the system to be in a position, such that if there is anything wrong with the overall process of forming beliefs, it will be alerted to it and respond appropriately. Otherwise, if there is nothing wrong, the system can accept the deliverances of its belief-forming processes by default, without the further requirement to provide any explicit positive reasons in their support. This is a form of epistemic responsibility that does not belong to any of the component parts but to the relevant system as a whole. The reason is that reliability does not arise on the basis of any component parts operating in isolation but instead on their ongoing interactivity, which, according to dynamical systems theory, belongs to the system as whole.[13]

For example, it is possible to use the above approach in order to explain how a subject might come to perceive the world on the basis of a Tactile Visual Substitution System (TVSS), while also holding fast to the idea that knowledge is belief that is true in virtue of

[12] Elsewhere (Palermos 2011, Palermos 2014*b*), it has been argued that both theories also put forward the same broad, common sense functionalist intuitions on what is required from a process to count as a cognitive ability. Briefly, both views state that the process must be (a) normal and reliable, (b) one of the agent's habits/dispositions and (c) integrated into the rest of the agent's cognitive character/system.

[13] Recently there has been a number of process reliabilist attempts to account for knowledge that is produced on the basis of epistemic artifacts or in a socially distributed fashion. For example Goldberg (2010) has put forward the extendedness hypothesis, Goldman (2014) has proposed social process reliabilism and Michaelian and Arango Muñioz (Michaelian 2014; Michaelian and Arango Muñioz forthcoming) have offered the alternative of distributed reliabilism. An anonymous referee therefore wonders why I here choose to focus on virtue reliabilism as opposed to process reliabilism. The main reason has to do with the paper's main claim that there is a way to design Social Machines such that their socio-epistemic properties can be viewed as the cognitive properties of a distributed cognitive system. In contrast to virtue reliabilism, traditional forms of process reliabilism (and presumably the above derivative accounts) explicitly state that cognition resides strictly within the head or organism of individual agents. Accordingly, it would be hard to motivate, on the basis of process reliabilism, the claim that TMSs or Social Machines are cognitive systems in Berners-Lee's sense. The contrast between the above process reliabilist accounts and virtue reliabilism, when it comes to explaining extended and distributed knowledge, becomes most evident with respect to the issue of epistemic responsibility. Michaelian and Muñoz write for example: Assignments of credit and responsibility for cognitive success and failure "presuppose a richer notion of agency: what we might refer to as *responsible* cognitive agency, where responsible cognitive agency requires cognitive agency, plus responsibility. There is no clear sense in which a TMS, for example, might be assigned responsibility for its cognitive success and failures, and it is in this sense we have suggested that extended and distributed memory systems do not qualify as cognitive agents." Process reliabilism therefore can't account for distributed epistemic responsibility, which is a reason to suggest that the relevant distributed system cannot qualify as a cognitive agent. In contrast, as the above indicates, virtue reliablism employs the notion of cognitive integration in order to not only account for epistemic responsibility but also explain how the resulting epistemic responsibility belongs to an extended or even distributed *cognitive* system as a whole. This renders virtue reliabilism a significantly more promising candidate for explaining how Social Machines can be designed in order to qualify as epistemic cognitive systems in themselves. Perhaps, however, a modified version of distributed process reliabilism along the lines suggested by (Michaelian 2014; Michaelian and Arango Muñioz forthcoming) could provide the resources for conceptualizing distributed agents as proper cognitive—*qua* responsible—agents, but such an approach would require departing from the more traditional accounts of process reliabilism that restrict cognition within the head of individual cognitive agents.

cognitive ability (i.e. the ability intuition on knowledge). A tactile visual substitution system is a mini video camera attached on a pair of glasses, which converts the visual input into tactile stimulation under the agent's tongue or her forehead. By moving around and on the basis of the associated sensorimotor contingencies,[14] blind patients quickly start perceiving shapes and objects and orienting themselves in space. Occasionally, they also offer reports of feeling as if they are *seeing* objects, indicating that they are enjoying phenomenal qualities very close to those of the original sense modality that is being substituted. In the light of DST, therefore, seeing through a TVSS qualifies as a case of cognitive extension, because it is a dynamical process that involves ongoing reciprocal interactions between the agent and the artifact. By moving around, the agent affects the input of the mini-video camera, which continuously affects the tactile stimulation she will receive on her tongue or forehead by the TVSS, which then continuously affects how she will move around and so on. Eventually, as the process unfolds, the coupled system of *the agent and her TVSS* is able to identify—that is, see—shapes and objects in space.

But what about *distributed* cognitive abilities? Can knowledge arise on the basis of collective cognitive abilities that emerge out of several individuals' mutual, socio-epistemic interactions? Indeed, there have already been several attempts to introduce this sort of collective knowledge within the literature. Think for example of true beliefs produced by scientific research teams. As several philosophers and ethnographers of science suggest, employing the framework of distributed cognition is the most promising way to analyze such collaboratively produced scientific knowledge (Giere 2002*a*; 2002*b*; 2006*;* 2007; Giere & Moffat 2003; Knorr-Cetina 1999; Nersessian et al. 2003*a*; 2003*b*; Neresessian 2005; 2006; Thagard 1993; 1994; 1997; Palermos 2015)

Similarly, consider the case of transactive memory systems (TMSs) (Wegner et al. 1985; Wegner 1986) within cognitive psychology. TMSs are groups of two or more individuals that collaboratively encode, store and retrieve information:

> 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

[14] For a recent review on TVSS, see Bach-y-Rita and Kercel (2003). 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 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.

> Okanogan. Rudy says, "Yes, in that area with all the fruit stands." Lulu finally makes the identification: Peachland (Wegner et al. 1985, p. 257).

As Wegner et 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, p. 257). Similarly, Barnier et al. (2008) suggest that TMSs always involve skillful interactive simultaneous coordination between their members, rendering them particularly good candidates for counting as distributed cognitive systems. In Wegner and his colleagues' words, the members' interaction "gives rise to a knowledge-acquiring, knowledge holding and knowledge-using system that is greater than the sum of its individual member systems" (1985, p. 256).

TMSs are currently at the forefront of several studies within philosophy of mind and cognitive science (Barnier et al., 2008; Sutton et al., 2010; Sutton, 2008; Wegner et al., 1985; Wegner 1986; Theiner et al., 2010). In the present context, such studies can be particularly useful for revealing some of the *practical* preconditions for building *cognitively integrated* TMSs and—by extension—distributed cognitive systems in general. As Wegner et al. note, "to build a transactive memory is to acquire a set of communication processes whereby two minds can work as one (Wegner et al. 1985, p. 263). The first step towards this objective 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—i.e., 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.

In return, this allows the members of the emerging TMS to take the second step, which is to grow the *differentiation* of their group: Couples typically begin a relationship by revealing information about themselves to each other. 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. Eventually, each member gets a sense of who the other members are, and thereby knows when it is appropriate to rely on the knowledge and expertise of the others and, conversely, when it is time to take action themselves.

Finally, once common knowledge and differentiation are in place, the dyad is ready for the third and final step towards the acquisition of a TMS—i.e., the formulation of an

integrated structure. According to Wegner et al., an *integrated structure* effectively comes about when and only when its members can efficiently take advantage of their common knowledge and differentiated structure so as to create new knowledge on the basis of effective communication feedback loops.[15]

Interestingly, similar things can also be said about scientific research teams. In order for the members of a research team to communicate with each other they must share some common knowledge or, in other words, what Kuhn (1962) calls a 'scientific paradigm' (briefly, a set of agreed-upon metaphysical assumptions, methodological rules, scientific theories, techniques for using equipment, etc.). Nevertheless, it wouldn't make sense for the members of a research team to have *everything* in common. In order to efficiently allocate the workload of performing a complex experiment or solving a research problem, members must also possess some expertise, which will provide their team with a differentiated structure. This differentiated structure must be part of the members' common knowledge so that everyone will know whom they need to communicate with when they face a problem they cannot solve on their own or when they think that they have stumbled upon an interesting finding that might be useful with respect to the overarching goal of the team. Once these practical prerequisites are in place, the members of the research team can start collaborating efficiently by mutually interacting with each other.

TMSs and certain scientific research teams can therefore act as exemplars whereby justification/epistemic responsibility emerges on the basis of the members' socio-epistemic interactions. These two distinct cases make apparent that in order to have a cognitively integrated distributed cognitive system—whose members are in a position to mutually interact with each other—the relevant group must possess some *common knowledge* and a *differentiated structure*.

In the following section, we shall see how these points can offer significant input to the efficient design of future distributed cognitive systems. Central to present purposes, however, is the fact that they reveal the following theoretical point: Distributed cognitive systems are not simply capable of manifesting just any higher-level cognitive process. Instead, provided that the right preconditions are in place, distributed cognitive systems are capable of manifesting two particularly advanced cognitive abilities—i.e., the abilities to acquire knowledge and justification. Asa big part of todays' social computing is oriented towards the

[15] The above points regarding the emergence of TMSs may create the impression that distributed cognitive systems need time to develop a sufficiently integrated structure. Nevertheless, any of the above preconditions to cognitive integration may in principle be very quickly realized within a group, provided that the appropriate socio-technical system is in place. It is true, however, that as time goes by, self-organising systems, such as distributed cognitive systems, can increase their efficiency by adapting to the conditions of the dynamic environments in which they operate. Currently, most distributed cognitive systems would need some time before achieving an appropriate degree of integration, but this does not mean that we could not possibly design socio-technical systems that would be very quick to achieve the required state or even start with it by default.

production, storage and dissemination of reliable information this is a possibility that is not easy to overstate.[16] As Halpin et al. (2010) note,

> A central challenge [for the future of the Web] will be to analyze the conditions under which users trust, by responding unreflectively and uncritically to the collectively maintained information retrieved from the Web. [...] From Wikipedia to Google, this ability to trust information on the Web is one of the most pressing problems facing the Web today [(O'Hara & Hall, 2008)], so a detailed philosophical analysis of this topic promises to have a significant impact on the practice of Web engineers designing these systems.

3. Social Machines

What then does the design of Social Machines require, in order to manifest higher-level cognitive properties such as knowledge and justification? And, are such socio-technical systems already available?

The recent advent of Wikis has played a catalytic role in attempting to build such biotechnologically hybrid systems. Wikis is a characteristically emerging technology of the Social Web, whose main purpose is to "support users in generating, editing and organizing content online", by providing "a suite of applications, services, technologies, formats, protocols and other resources, all united in their attempt to both foster and support social interactions" (Smart and Shadbolt, 2014).

> A Wiki allows for non-linear, evolving, complex and networked texts, argument and interaction and is ideal for an environment where multiple persons are contributing to a project because a user can read, add or modify content to a Wiki page. Users can track changes made to a Wiki page and other users can be automatically notified by email of such changes" (Black et al., 2007, pp. 246-7).

Wikis allow everyone to be "active producers of expertise rather than passive consumers of information" (very much in the spirit of Berners-Lee's vision for the future of the Web) and they can be thought of as "species of the genus known as collaborative editing software" (Noveck, 2007).

Wikis therefore are a fitting platform for building Social Machines, because they can provide the means to instantiate the mutual interactions that are necessary for the emergence of distributed cognitive systems. Nevertheless, having the means to interact densely does not guarantee that the resulting social system is going to be well-integrated. To illustrate this, while keeping in mind the previous section's points on *cognitive integration*, it will be helpful to focus on the most famous and impressive Social Machine to date—namely Wikipedia.

Wikipedia is an open, collaborative Wiki that anyone can edit either anonymously or eponymously. Even though it does not require any credentials or expertise on the part of the editors, the sheer number of the people who take interest in participating and correcting

[16] For more discussions of TMSs in the context of Web science see (Sparrow and Chatman 2013) and (Sparrow, Liu and Wegner 2011).

mistakes has made Wikipedia as reliable as more conventional encyclopedias.[17] Specifically, the huge number of participants in combination with the fact that Wikipedia is easily searchable via search engines allows for its content to continuously receive a degree of scrutiny that no single editor could ever provide on his or her own, irrespective of his or her level of expertise. In the form of a slogan: Wikipedia uses the "power of the many eyes" (Noveck, 2007): The huge amount of attention makes it possible for everyone to participate, because, even if people post inaccuracies, there is always someone else to deal with them. Wikipedia, in other words, generates knowledge not by maximizing the probability of true posts, but by minimizing the probability that falsehoods will survive.

This unorthodox strategy of 'error minimization', as mirrored by Wikipedia's policy of completely free editability, is a crucial part of its success story: Wikipedia—originally developed from another encyclopedia project (viz., Nupedia)—is perhaps the best example of the snowball effect within Web science. Launched on January 15, 2001, by September 25 of the same year Wikipedia had more than 13,000 articles—a number that has now increased to 4,6 million entries.[18]

Impressive as this number of articles may sound, Wikipedia's success can be best represented in terms of the number of its active contributors (i.e., the editors who after editing Wikipedia once, keep coming back with equal or more than 5 entries per month). Arguably, active contributors are Wikipedia's driving force and, as a recent study demonstrates (Halfacker et al., 2012), between 2001 and 2007, their number grew exponentially, starting from a few hundreds and reaching its peak at 56,400, in March 2007.

These are impressive numbers that secure Wikipedia a prominent position in the history of Web science. Yet Wikipedia has its Achilles' heel. As the same study demonstrates, since 2007, Wikipedia has been facing a particularly worrisome, steady decline in the number of its active contributors. Halfacker et al. (2012) argue that the problem begun around 2006, when Wikipedia's exponential growth made the Social Machine particularly attractive to vandals. In response to this problem, Wikipedia introduced automated bots (e.g., ClueBot NG) and semi-automated tools (e.g., Twinkle, rollback and Huggle) that "dramatically increased the efficiency of Wikipedia's quality control system." Ironically, however, those very tools are also the cause behind the decline in the number of Wikipedia's active contributors.

Algorithmic tools tend to ban not only acts of vandalism but also entries, which even though do not comply with Wikipedia's standards, they have been written *in good faith*, such that after appropriate feedback and revision could be "massaged" into form and turned into valuable contributions. Allowing algorithmic tools to reject such "'unwanted' but not

[17] At least with respect to a wide range of scientific topics. See (Giles, 2005). See also encyclopedia Britannica's response (http://corporate.britannica.com/britannica_nature_response.pdf) and *Nature's* counter response (http://www.nature.com/press_releases/Britannica_response.pdf).
[18] All numbers refer to English Wikipedia alone.

intentionally damaging contributions" almost by default is itself an obvious problem for Wikipedia, which in effect misses out on some potentially good contributions. But the main problem with this strategy goes deeper: Fully automated tools make the provision of meaningful editorial feedback impossible or, if possible (as is the case with some semi-automated tools), they leave it entirely on the users' discretion, who unfortunately, most times, do not bother.

According to Halfacker et al. (2012), this is at the very heart of the problem with the declining number of Wikipedia's active contributors: Good-faith newcomers do not appreciate having their entries rejected or reverted without any explanation about what went wrong and how they could make them better. This makes them feel as if they lost their time, which, in most cases, is sufficient for putting them off from contributing ever again. In order for newcomers to keep coming back, such that they can become active contributors, they must instead be able to receive meaningful feedback by interacting, not just with impersonal, automated bots and semi-automated tools, but with the rest of the Wikipedia community.

This is the problem that Wikipedia is currently facing. Before attempting to propose a solution, it is interesting to ask whether Wikipedia could have actually predicted and prevented this problem all along. The answer, strictly speaking, could in fact be positive. As Berners-Lee suggests in his initial definition of a Social Machine, the machine aspect of the Social Machine is supposed to be doing only the administrative work, whereas the creative work should be down to the human components alone. Clearly, though, deciding whether the quality of an entry is sufficiently high is a creative task, and as such it'd better not be left to the machine aspect of the social machine alone.

Nevertheless, as Smart and Shadbolt (2014) point out, the dividing lines of Berners-Lee's definition may not be as clear-cut. Specifically, with respect to the completion of creative tasks, it may not always be possible to distinguish between human and machine contributions—especially within the context of the extended cognition hypothesis:

> It is not always clear that human creativity is something that can (easily) be accomplished in the absence of some form of bio-technological or bio-artifactual entanglement. Many cases of human creativity could be said to be ones in which extra-organismic resources are playing some form of representationally and/or computationally significant role. In the process of writing an academic paper, for example, we often witness a rich series of brain-body-world interactions that serve to productively constrain the nature of what gets written (Clark, 1997, p. 207).

Even though there is no room for disagreement with Smart and Shadbolt on this, notice that Berners-Lee's distinction is still not entirely void of meaning. As Smart and Shadbolt note, creative tasks will many times involve artefacts in one way or the other, such that it would be naïve to disallow the involvement of machines to anything that might be potentially classed as creative activity. But a productive and perhaps more helpful way to interpret Berners-Lee's

remarks is to keep in mind that, at the very least, no creative task should ever be handled in its entirety by a Social Machine's machine aspect *alone*. Wikipedia seems to have committed precisely this mistake.

In order to solve its problem, therefore, Wikipedia must find a way to keep the quality of its massive content high without just using automated bots or semi-automated tools.[19] Up until its exponential growth, Wikipedia was in a position to achieve this, merely by relying on its policy of free editability and the sheer number of Wikipedians. After 2007, and even as early as 2006, however, this was not possible anymore. The vast amount of entries and the increased vandalism attacks created a labour squeeze that required new measures. But what can those measures be, if not the introduction of vigilant machines? The response that cognitive science and epistemology seem to suggest is that Wikipedia needs to *cognitively integrate*. It needs to structure itself so that the contributors can densely interact with each other.

While there are several ways to achieve this, the above suggests that Wikipedia (or any other Social Machine for that matter) can also use the help of some general guidelines. Specifically, Wikipedia could attempt to take advantage of the two properties—i.e., *common knowledge* and *differentiated structure*—that were previously noted to be *practically* necessary for having a well integrated, distributed cognitive system. Moreover, Wikipedia should also focus on the process *of self-organisation*, which, according to DST, is intimately related to the existence of *feedback loops* between the components of the relevant system.

There may again be several ways in which Wikipedia could implement these properties in its biotechnologically hybrid algorithm, but, to offer a concrete example, a promising way for doing so is to take the following two steps. The first step is to ask every newcomer and existing contributor to register some areas of expertise—though, of course, without the provision of any credentials whatsoever (otherwise Wikipedia would go against its own policy of completely free editability). Wikipedia can then use this information in order to more efficiently allocate the workload of editing Wikipedia, by sending notifications of a new entry to only those editors who have reported to possess the relevant expertise.

Note that a potential problem with the above step is that it sounds arbitrary, in the sense that it does not guarantee that competent users are going to edit the relevant entries. This is precisely what the second step is meant to ensure: Throughout the whole process, Wikipedia can continuously keep monitoring how many changes a given contributor's edits

[19] Contrary to what I claim above it may be thought that Wikipedia *can* keep its quality high merely on the basis of bots. But this is overly optimistic as to what bots can do and underestimates the difficulty and degree of understanding what is required for completing the task. For example, while a bot can detect that an entry lacks a sufficient amount of references, it could not possibly assess the reliability and the appropriateness of the references provided. Indicatively see Metz's article 'Wikipedia Deploys AI to Expand its Ranks of Human Editors', published at *WIRED*: http://www.wired.com/2015/12/wikipedia-is-using-ai-to-expand-the-ranks-of-human-editors/?mbid=social_fb

(on any given domain of expertise) undergo over time and if their number is considerably high, recall that editor's status of expertise on the relevant problematic domain (see figure 1).

Both steps are designed to promote Wikipedia's *cognitive integration.* From the point of view of epistemology, they can allow Wikipedia to operate as an integrated, epistemically responsible agent in itself. By implementing the above elements in its algorithm, Wikipedia won't just manage to allow the right contributors to densely and meaningfully interact with each other by providing comments to each other's entries. It will also manage—as the feedback loop of figure 1 indicates—to epistemically *self-regulate*. Specifically, it will make it possible for the right contributors to keep monitoring each other, thereby having a direct effect on each other's status of expertise and, as a result, on the kinds of editing notifications they will receive in the future (based on how many changes one's own edits—in any given domain of expertise—undergo over time). In this way if there are any mistakes, vandalism, falsehoods or vague claims posted online, they will be swiftly spotted and immediately removed—otherwise, the remaining entries can be considered to amount to knowledge, by default.

In other words, if we recall how epistemology may account for individual knowledge and epistemic responsibility/justification, the above two steps will allow Wikipedia to efficiently generate knowledge in functionally the same way that we generate knowledge within our own heads. By allowing the different components of the overall integrated cognitive system to keep monitoring each other, it will be in a position to swiftly spot any shortcomings and deal with them in an appropriate way. Otherwise—if nothing's wrong—the overall system will be able to count its propositional attitudes as knowledge by default.

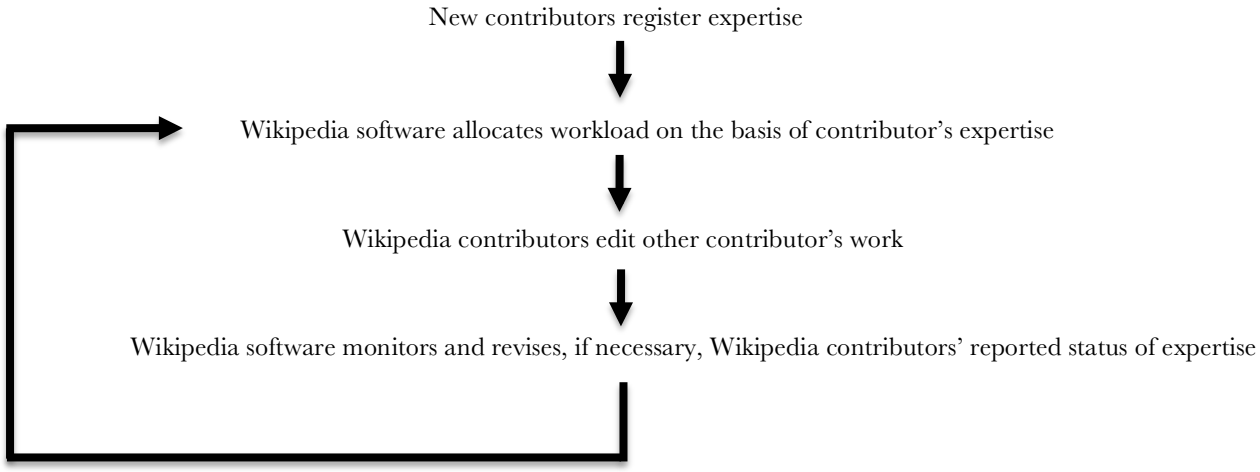

Figure 1. A workflow that could facilitate Wikipedia's cognitive integration.

Similarly, the above two steps also contribute to *cognitive integration* from the point of view of active externalism within philosophy of mind and cognitive science. Specifically, they are meant to facilitate the underlying components' reciprocal interactions, which are necessary for the emergence of the overall system's collective properties: As the workflow of figure 1 indicates, every contributor will have the chance, on the basis of his/her registered expertise, to affect the others. At the same time, by being monitored and having their expertise recalled as and if required, every contributor will be potentially affected by the rest, who will determine what he or she can bring back into the group. In this way, Wikipedia will manage to *self-organise* by restricting the behaviour of its members in a way that will allow them to achieve a stable configuration that will efficiently give rise to Wikipedia's collective properties—properties, such as its overall reliability, which do not stem from any single individual's expertise, but from the operation of the system as a whole.

Specifically, when organised in the way indicated above, the reliability of Wikipedia cannot be reduced to the reliability of any of its individual components—recall that Wikipedia has always operated on the basis of a completely free editability policy, whereby anyone can post and amend entries in an anonymous way, without providing any form of credentials. Instead, its reliability derives from the actual and potential interactions of its biotechnologically hybrid algorithm, which comprises of a set of administrative instructions provided primarily by its mechanistic aspects along with the on-going creative responses provided primarily by its active contributors.[20] According to the DST approach to active externalism, this form of reliability is an emergent, distributed property that belongs to Wikipedia as an overall, integrated system and cannot be accounted for in terms of the reliability of its underlying mechanistic and organismic components. Its reliability arises, instead, out of the actual and potential synergetic interactions of all components—organismic and mechanistic ones alike—at the same time.

In this way, seen from the point of view of epistemology and philosophy of mind and cognitive science, Wikipedia may qualify as an emergent, epistemically self-regulating and

[20] These points can help disambiguate what may constitute the distributed cognitive system of Wikipedia. An anonymous referee worries that we may have to accept that Wikipedia and all of its creators and users across time constitute a single distributed cognitive system. This seems to be rather implausible and unlike the paradigm distributed cognitive system in the literature—i.e., the crew-members of a ship as it performs near land manoeuvres (Hutchins 1995). In contrast to the navy ship, where all members are involved in the on-going process of navigating the ship, not all of Wikipedia's creators and users are involved in Wikipedia's on-going processes. To address this worry, we need to keep in mind that systems are individuated on the basis of the processes we are interested in (for more details on system individuation see (Palermos 2014a) and (Palermos forthcoming)). In order to figure out what sort of system Wikipedia really is we need to look at what components are currently giving rise to the Wikipedia processes/properties that we are interested in, such as its overall reliability. Whatever components currently contribute to this process in a constitutive fashion (by mutually interacting with each other) can be considered as proper parts of Wikipedia's distributed cognitive system. As indicated in the main text, this will primarily consist by the socio-techinical system that comprises of Wikipedia's algorithm and its active contributors. As time goes by, different parts of this socio-technical system might be substituted with others (e.g., active contributors may come and go), but whatever (biological and technological) components give rise to Wikipedia's reliability (by mutually interacting with each other) at any given time, they can be considered as constitutive parts of Wikipedia's cognitive system at that time.

cognitively self-organising system in its own right, which is capable of producing knowledge and justification in a way that is functionally similar to the way knowledge arises within individual heads.[21]

4. The Future of Social Machines

The foregoing make apparent the potential impact of epistemology and cognitive science for the efficient design of Social Machines, with Wikipedia being only one example among potentially a multitude of different cases. But what about the overarching question of what may qualify as a Social Machine?

Previously we noted that building on Berners-Lee's initial suggestion that Social Machines are Web-driven processes designed to give rise to an overall mechanism that can work as unified brain, Smart and Shadbolt have suggested the following definition: Social Machines are "web-based socio technical systems in which the human and technological elements play the role of participant machinery with respect to the mechanistic realization of system-level processes." (2014). While this is an improvement, we noted that this definition is bound to be incomplete in the absence of a clear understanding of what may count as 'participant machinery' and 'system-level processes.' In order to clarify these two notions, we suggested that we need to focus on the concept of *cognitive integration*—i.e., we need to understand when some component may count as proper part of a cognitive system.

With the foregoing in mind, it is now possible to complete Smart and Shadbolt's definition. Specifically, on the basis of the previous discussion on DST, we can here claim that 'system-level processes' (or properties) are emergent regularities in the behaviour of the components of integrated systems. Moreover, such systems count as integrated precisely because their components are mutually interacting with each other and can thereby count as 'participant machinery'. It is this *mutual interdependence* between component processes that renders the otherwise seemingly distinct parts into 'participant machinery' of the larger, integrated system's working.

Overall then, Smart and Shadbolt are correct to note that the "term [(social machines)] is used to draw attention to Web-based systems that feature some degree of active human participation, and it is this notion of active human participation that seems to be critically important to what makes something a social machine." (2014). However, active externalism and epistemology further point out that it is not just any active participation that will do. Instead, humans must *both affect and be affected* by each other as well as the machine

[21] An anonymous referee raises the question of how Wikipedia could implement the suggested change in the way it operates. As Halfacker et al. (2012) report, Wikipedia is currently working on the basis of enforced formal policies. Most of them are decided by the board and stuff members of Wikimedia Foundation. A change in Wikipedia's software, such as the one suggested here, would also require to go through Wikimedia's system administrators https://meta.wikimedia.org/wiki/System_administrators.

components of the overall system, such that their own actions may be partly determined by the system and *vice versa*. This two-way interaction between individual components and the socio-technical system is crucial, because, it is precisely this feature of Social Machines that distinguishes them from other kinds of social computing that also involve active human participation (think for example crowdsourcing (Doan et al., 2011)).

To see how the above may translate in practice it is helpful to use Chi's (2008) breakdown of the Social Web collaboration spectrum into 'light', 'middle' and 'heavy' Social Web. According to Chi, on the left side of the spectrum we have what he calls 'Lightweight Social Web' which includes *crowdsourcing* software such as digg, reddit, and delicious, where users can only participate actively by means of a rather informationally restricted way; for instance, mainly in terms of votes.[22] Then, in the middle, we have *collective information structures*, such as YouTube, Facebook, and Twitter where users can both post and vote for content, but where the creation of content is still largely an individual act.

From there, Chi goes straight to the right end of the spectrum, which is supposed to be populated by Social Machines alone. Recent developments, however, suggest that an additional category may be introduced. Specifically, even though initially absent from Chi's original breakdown, it might be helpful to introduce the additional category of 'junior heavy weight Social Web' in order to appropriately categorise *open innovation* software, such as Galaxy Zoo and Planet Hunters.[23] The distinctive feature of such software is that it provides a multitude of tools that allow users to contribute in a range of ways that clearly surpasses the opportunities for interaction provided by collective information structures.

Nevertheless, just like in the case of 'Middle Weight Social Web' and unlike in the case of Social Machines, the content of most open innovation structures is still primarily created by individuals working on their own.[24] Only in the far right, 'heavy weight' end of the spectrum—which includes collaborative co-creation software and Social Machines like Wikipedia—is the creation of content a process that is *distinctively social*.[25]

If the above is correct, then we seem to have a straightforward way for classifying the main kinds of social computing, while also accentuating the distinctive nature of Social Machines. It is important to also ask, however, whether computer scientists are likely to accept such a classification—and if not, shouldn't such a rejection by the practitioners of the

[22] Digg and reddit are entertainment, social news networking services. Their members can submit content, such as text posts or direct links. Registered users can then vote submissions up or down to organize the posts and determine their position on the sites' pages.
[23] Planet Hunters is another citizen science project that enlists members of the public to assist with finding planets.
[24] To say that some cases of the available open innovation software do not qualify as Social Machines is not to deny that they are social systems. Moreover, some open innovation software may occasionally qualify as Social Machines, provided that the criterion of mutual interactions between the contributing members is satisfied. One such a case may well be the discovery of a new type of galaxies known as Green Pea galaxies, which started as a discussion in the internet forums of Galaxy Zoo, with he name "Give peas a chance", and in which various green objects were posted.
[25] As Shadbolt et al. (2013) put it, " the key difference between social machines and open innovation is at the level of interaction between the social and the machine-driven processing components" (3).

field count as a *reductio* against the present argument (simply put, that Social Machines are meant to be Web-driven distributed cognitive systems)?

As noted in the beginning, computer scientists seem to indiscriminately apply the term 'Social Machine' to a large number of Web platforms such as Facebook, mySpace, Twitter, Galaxy Zoo, Wikipedia, reddit, digg, etc. Therefore, an objection could be advanced to the effect that any account that restricts the term 'Social Machines' to only those online platforms that allow their users to mutually interact with each other along with the platform itself should be discarded, because it runs against the well-established use of the term 'Social Machine' within Web science.

There are several things to be said in response. First, research on Social Machines is at a very early stage such that the way the term is being used may count as anything but 'well established.' Moreover, this point can be further reinforced if we consider that the above candidates for qualifying as Social Machines do not have anything in common other than being instances of social computing.

Of course, one could further claim that 'Social Machine' is not supposed to be a proper name but rather an umbrella term that may simply act as a synonym of the term 'social computing.' However, such a move would render the term 'Social Machines' redundant and it would blur the subject matter of several research projects that are specifically dedicated to the study of Social Machines, rather than the study of social computing in general. Moreover, in general and as Chi's breakdown of the Social Web collaboration spectrum indicates, computer scientists are particularly keen on establishing ontologies that could further guide their research, such that one would expect that a sharp definition of the term 'Social Machine' would be welcome.

Perhaps then, if one were to reject the present approach, it would have to be with regards to the particular definition of 'Social Machines' on offer. Nevertheless, it should be kept in mind that the offered definition of 'Social Machines' is anything but arbitrary. On the contrary, it is a definition that is specifically designed to give flesh to Berners-Lee's programmatic claim that Social Machines refer to "interconnected groups of people acting as if they shared a larger intuitive brain." This is a bold claim indeed, but it is also one that philosophy of cognitive science and epistemology can shed significant light on.

Skepticism may of course persist about whether this is a goal that computer science should aspire to. Perhaps one may think that we do not need to go all the way with Berners-Lee's initial idea and perhaps less elaborate socio-technical systems could qualify as Social Machines. But wouldn't that ultimately amount to a mere terminological point? After all, according to epistemology and philosophy of cognitive science—and especially their focus on the process of cognitive integration and the phenomenon of distributed cognition—there seems to be a clear distinction between mere social systems and ones that allow their members

to act as if they were part of a unified distributed cognitive system. Should we invent a new term for such promising systems—perhaps 'Social Machines 2.0'—while leaving the older term undefined for computer scientists to use however they see fit? This may well end up to be the case, but if so, care should be taken so that such a move would not obstruct the transforming potential of Berners-Lee's initial idea.

Acknowledgements

I am thankful to Paul Smart for his feedback to a previous draft of the paper. I am also thankful to the audiences of the Computational Social Science Satellite Workshop of the European Conference on Complex Systems, (Lucca, Italy, 2014), the Connected Life Conference (Oxford Internet Institute, Oxford University, 2015) and the Extended Knowledge Impact Event (University of Edinburgh, January 2016), where previous versions of the paper were presented and discussed. The paper was produced as part of the AHRC-funded 'Extended Knowledge' research project (AH/J011908/1), which was hosted at Edinburgh's *Eidyn* Research Centre.

References

Adams, F.,&Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14(1), 43–64.

Adams, F., & Aizawa, K. (2010). *The bounds of cognition* (1st ed.). Malden, MA: Wiley-Blackwell.

Bach-y-Rita, P., & Kercel, S. W. (2003). Sensory substitution and the human-machine interface. *Trends in Cognitive Science,* 7(12), 541–6.

Barnier, A. J., Sutton, J., Harris, C. B., &Wilson, R. A. (2008). A conceptual and empirical framework for the social distribution of cognition: The case of memory. *Cognitive Systems Research*, 9(1–2), 33–51

Berners-Lee, T., Fischetti, M., & Foreword By-Dertouzos, M. L. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. Harper Information.

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., & Weitzner, D. J. (2006). A framework for web science. *Foundations and trends in Web Science*, *1*(1), 1-130.

Black, P., Delaney, H., & Fitzgerald, B. (2007). Legal issues for wikis: The challenge of user-generated and peer-produces knowledge, content and culture. *eLaw J.*, *14*, 245

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Harvard University Press.

Braddon-Mitchell, D., & Jackson, F. (2006). *Philosophy of mind and cognition: An introduction* (2nd ed.). Malden, MA: Wiley-Blackwell

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge: MIT press.

Chi, E. H. (2008). The Social Web: research and opportunity. *IEEE Computer*. 41 (9): 88-91.

Chisholm, R. M. (1977). *Theory of knowledge*. Prentice-Hall.

Clark, A. (2008). *Supersizing the mind.* Oxford University Press.

Clark, A. (2007). Curing cognitive hiccups: A defense of the extended mind. *The Journal of Philosophy*, 104, 163–192.

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. Cambridge, Massachusetts, USA: MIT Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Ananlysis*, 58(1), 7–19.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM*, *54*(4), 86-96.

Froese, T., Gershenson, C., & Rosenblueth, D.A. (2013). The dynamically extended mind. Retrieved http://arxiv.org/abs/1305.1958.

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

Giere, R. (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.

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

Giere, R. (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.

Giles, J. (2005). Internet encyclopaedias go head to head, *Nature,* 438 (7070), p. 900-901.

Goldman, Alvin and Beddor, Bob, "Reliabilist Epistemology", *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2015/entries/reliabilism/>.

Goldman, A. I. (2014). Social Process Reliabilism. In *Essays in Collective Epistemology*, Lackey, J. (ed.), Oxford University Press.

Goldberg, S. (2010). *Relying on others: an essay in epistemology*. Oxford: Oxford University Press.

Greco, J. (1999). Agent reliabilism. In James Tomberlin (Ed.), *Philosophical perspectives* 13: Epistemology. (pp. 273–296). Atascadero, CA: Ridgeview Press

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

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

Greco, J. (2010). *Achieving knowledge: A virtue-theoretic account of epistemic normativity*. Cambridge: Cambridge University Press.

Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2012). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, DOI: 0002764212469365.

Halpin, H., Clark, A., & Wheeler, M. (2010). Towards a philosophy of the web: representation, enaction, collective intelligence. In: *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*.

Halpin, H. (2013). Does the web extend the mind?. In *Proceedings of the 5th annual ACM web science conference* (pp. 139-147). ACM.

Hendler, J., & Berners-Lee, T. (2010). From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. *Artificial Intelligence*, *174*(2), 156-161.

Heylighen, F.,Heath,M.,&Van Overwalle, F. (2007). The emergence of distributed cognition: Aconceptual framework. In *Proceedings of collective intentionality IV* (2004), Vol. IV. University of Siena.

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

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

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: The University of Chicago Press.

Menary, R. (2007). *Cognitive integration: Mind and cognition unbound*. Basingstoke: Palgrave McMillan.

Michaelian, K. (2014). JFGI: From distributed cognition to distributed reliabilism. *Philosophical Issues*, *24*(1), 314-346.

Michaelian, K., and S. Arango-Muñoz. (forthcoming). Collaborative memory knowlege: A distributed reliabilist perspective. Collaborative Remembering: How Remembering with Others Influences Memory. Eds. M. Meade, A. Barnier, P. Van Bergen, C. Harris, and J. Sutton. Oxford University Press.

Nersessian, N. J., Newstetter, W. C., Kurz-Milcke, E. & Davies, J. (2003). A Mixed-method Approach to Studying Distributed Cognition in Evolving Environments.*Proceeedings of the International Conference on Learning Sciences.* pp. 307 - 314.

Nersessian, N. J., Kurz-Milcke, E., Newstetter, W. C., & Davies, J. (2003).Research laboratories as evolving distributed cognitive systems. *Proceedings of The 25th Annual Conference of the Cognitive Science Society.* pp.857-862.

Nersessian, N. J. (2005). Interpreting scientific and engineering practices: Integrating the cognitive, social, and cultural dimensions. In *Scientific and Technological Thinking*, M. Gorman, R. Tweney, D. Gooding, & A. Kincannon, eds. (Erlbaum). pp. 17-56.

Nersessian, N. J. (2006). The Cognitive-Cultural Systems of the Research Laboratory. Organization Studies, 27(1), pp. 125-145

Noë, A. (2004). *Action in Perception.* Cambridge, MA:MIT Press.

Noveck, B. S. (2007). Wikipedia and the Future of Legal Education. *J. Legal Educ.*, *57*, 3.

O'Hara, K. and Hall, W. (2008) Trust on the web: Some web science research challenges. *UoC Papers: E-Journal on the Knowledge Society*, (7).

Palermos, S. O. (2011). Belief-forming processes, extended. *Review of philosophy and psychology*, *2*(4), 741-765.

Palermos, S. O. (2014*a*). Loops, constitution, and cognitive extension. *Cognitive systems research*, *27*, 25-41.

Palermos, S. O. (2014*b*). Knowledge and cognitive integration. *Synthese*,*191*(8), 1931-1951.

Palermos, S. O. (2015). Active externalism, virtue reliabilism and scientific knowledge. *Synthese*, *192*(9), 2955-2986.

Palermos, S. O. (forthcoming). The Dynamics of Group Cognition. *Minds and Machines.*

Papineau, David (2015). Naturalism. *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2015/entries/naturalism/>.

Plantinga, A. (1993a). Warrant and proper function. New York: Oxford University Press.

Plantinga, A. (1993b). Warrant: The current debate. Oxford: Oxford University Press.

Pritchard, D. H. (2009). Knowledge, London: Palgrave Macmillan.

Pritchard, D. H. (2010a). 'Knowledge and Understanding', in A. Haddock, A. Millar & D. H. Pritchard, The Nature and Value of Knowledge: Three Investigations, Oxford: Oxford University Press.

Pritchard, D. H. (2010b). 'Anti-Luck Virtue Epistemology', manuscript, available at http://www.philosophy.ed.ac.uk/people/full-academic/duncan-pritchard.html.

Pritchard, D. (2010c). Cognitive ability and the extended cognition thesis. Synthese, 175, 133–151.

Rowlands, M. (1999). *The body in mind: Understanding cognitive processes*. New York: Cambridge University Press.

Rowlands, M. (2009). 'Extended Cognition and the Mark of the Cognitive'. *Philosophical Psychology*, 22(1); 1-19.

Rupert, R. D. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, 101(8), 389–428.

Rupert, R. D. (2009). *Cognitive systems and the extended mind* (1st ed.). Oxford: OUP USA.

Shadbolt, N. R., Smith, D. A., Simperl, E., Van Kleek, M., Yang, Y., & Hall, W. (2013). Towards a classification framework for social machines. In *Proceedings of the 22nd*

international conference on World Wide Web companion (pp. 905-912). International World Wide Web Conferences Steering Committee.

Shani, I. (2013). Making it mental: In search for the golden mean of the extended cognition controversy. *Phenomenology and the Cognitive Sciences*, 12(1), 1–26.

Smart, P. R. (2012). The Web-Extended Mind. *Metaphilosophy*, *43*(4), 446-463.

Smart, P. R. (2014). Embodiment, Cognition and the World Wide Web. In L. Shapiro (Ed.), *The Routlege Handbook of Embodied Cognition.* Routlege, New York.

Smart, P. R., and Shadbolt, N. R (2014) Social Machines. In, Khosrow-Pour, Mehdi (ed.) *Encyclopedia of Information Science and Technology.* Hershey, Pennsylvania, USA, IGI Global, 6855-6862.

Smart, P., Heersmink, R. & Clowes, R. (forthcoming). The Cognitive Ecology of the Internet. In S. Cowley & F. Vallée-Tourangeau (eds.), *Cognition Beyond the Brain: Computation, Interactivity and Human Artifice* (2nd ed.). Dordrecht: Springer.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*(6043), 776-778.

Sparrow, B., & Chatman, L. (2013). Social Cognition in the Internet Age: Same As It Ever Was?. *Psychological Inquiry*, *24*(4), 273-292.

Steup, M. (1999). A Defense of Internalism. In *The Theory of Knowledge: Classical and Contemporary Readings, 2nd edition*. Wadsworth Publishing

Sutton, J. (2008). Between individual and collective memory: Coordination, interaction, distribution. *Social Research,* 75(1), 23–48.

Sutton, J., Harris, C. B., Keil, P. G., & Barnier, A. J. (2010). The psychology of memory, extended cognition, and socially distributed remembering. Phenomenology and the Cognitive Sciences, 9(4), 521–560. doi:10.1007/s11097-010-9182-y

Thagard, P. (1993). Societies of minds: Science as distributed computing. *Studies in History and Philosophy of Science Part A*, *24*(1), 49–67. doi:10.1016/0039-3681(93)90024-E

Thagard, P. (1994). Mind, Society, and the Growth of Knowledge. *Philosophy of Science*, *61*(4), 629–645.

Thagard, P. (1997). Collaborative Knowledge. *Noûs*, *31*(2), 242–261. doi:10.1111/0029-4624.00044

Theiner, G., Allen, C., & Goldstone, R. (2010). Recognizing group cognition. *Cognitive Systems Research*, 11(4), 378–395.

Theiner, G., & O'Connor, T. (2010). The emergence of group cognition. In A. Corradini & T. O'Connor (Eds.), *Emergence in science and philosophy* (pp. 6–78). London: Routledge.

Tollefsen, D. (2006). From extendedmind to collective mind. *Cognitive Systems Research*, 7(2–3), 140–150.

Tollefsen, D., & Dale, R. (2011). Naturalizing joint action: A process-based approach. *Philosophical Psychology*, 25, 385–407.

Wegner, D., 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.

Wegner, D. (1986). Transactive Memory: A Contemporary Analysis of the Group Mind. *Theories of Group Behavior*. Eds. B. Mullen and G. R. Goethals. New York: Springer-Verlag.

Wilson, R. A. (2000). The mind beyond itself. InD. Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* (pp. 31–52). New York: Oxford University Press.

Wilson, R. (2004). *Boundaries of the mind: The individual in the fragile sciences: Cognition*. New York: Cambridge University Press.

Wilson, R. (2005). Collective memory, group minds, and the extended mind thesis. *Cognitive Processing*, 6(4), 227–236.