RETHINKING MACHINES: ARTIFICIAL INTELLIGENCE BEYOND THE PHILOSOPHY OF MIND

BY

DANIEL JOSE ESTRADA

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Philosophy
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Jonathan Waskan, Chair
Associate Professor Daniel Korman
Emeritus Professor Steven Wagner
Associate Professor Andrew Arana

Abstract

Recent philosophy of mind has increasingly focused on the role of technology in shaping, influencing, and extending our mental faculties. Technology extends the mind in two basic ways: through the creative design of artifacts and the purposive use of instruments. If the meaningful activity of technological artifacts were exhaustively described in these mind-dependent terms, then a philosophy of technology would depend entirely on our theory of mind. In this dissertation, I argue that a mind-dependent approach to technology is mistaken. Instead, some machines are best understood as independent participants in their own right, contributing to and augmenting a variety of social practices as active, though often unrecognized, community members.

Beginning with Turing's call for "fair play for machines", I trace an argument concerning the social autonomy of nonhuman agents through the artificial intelligence debates of the 20th century. I'll argue that undue focus on the mind has obscured the force of Turing's proposal, leaving the debates in an unfortunate stalemate. I will then examine a network theoretic alternative to the study of multi-agent complex systems that can avoid anthropocentric, mind-dependent ways of framing human-machine interactions. I argue that this approach allows for both scientific and philosophical treatment of large and complicated sociotechnical systems, and suggests novel methods for designing, managing, and maintaining such systems. Rethinking machines in mind-independent terms will illuminate the nature, scope, and evolution of our social and technological practices, and will help clarify the relationships between minds, machines, and the environments we share.

To Internet

Acknowledgements

This dissertation was a collective effort. The students and faculty at the University of Illinois, Illinois State University, and Johns Hopkins' Center for Talented Youth provided important material and intellectual support for these ideas. My mom, dad, and the rest of my family have never failed to be supportive of my eccentricities; responsibility falls squarely on their shoulders. Kyle Broom, Alex Spector, Kirk and Meghan Fisher, Joe Swenson, Shawn Dingle, Harmony Neal, Justin Remhof, Lally Gartel, Patrick O'Donnell, and Todd Kukla have each helped to nurture these ideas on late nights around warm fires, and the results bear the mark of their distinctive hands at work. Jon Lawhead deserves special recognition for being a lightning rod to my negative charge. Becca Spizzirri deserves a castle, in a park, in the rain.

New York, 2014

Table of Contents

Chapter 1: Introduction: Rethinking Machines 1

Chapter 2: Artificial Intelligence beyond the Philosophy of Mind 9

Chapter 3: Lady Lovelace and the Autonomy of Machines 53

Chapter 4: The Organization of Technological Networks 86

Chapter 5: Gaming the Attention Economy 113

References 136

Chapter 1

Introduction: Rethinking Machines

1.1 A prelude on machine participants

> "…but basically machines were not self-moving, self-designing, autonomous. They could not achieve man's dream, only mock it. They were not man, an author to himself, but only a caricature of that masculinist reproductive dream. To think they were otherwise was paranoid. Now we are not so sure. Late twentieth-century machines have made thoroughly ambiguous the difference between natural and artificial, mind and body, self-developing and externally designed, and many other distinctions that used to apply to organisms and machines. Our machines are disturbingly lively, and we ourselves frighteningly inert." - Donna Haraway[1]

I have a complicated relationship with my phone. No other object occupies as much of my time and focused attention. The other objects I use regularly, like my keys and wallet, usually just sit idly on the dresser when I'm home; but my phone stays within arms reach all day, whether I'm awake or asleep. After my glasses, my phone is the first thing I grab for in the morning and the last thing I put down at night; but unlike my glasses, my phone is a recurring object of attention throughout the time between. The computer in my office is where I do most of my work and play, but the computer stays in my office if I'm doing anything else, whereas my phone always stays with me. When in use, my phone occupies some of my most significant cognitive resources (my hands and my visual field) to the exclusion of just about everything else-- a risky proposition if I'm doing anything dangerous like driving a vehicle. My phone is a paradigm case of what Borgmann (1987) called a "focal thing": it provides a "center of orientation"[2] around which my daily habits and practices revolve.

And I'm not alone. According to a recent Pew survey, of the 90% of Americans that have cell phones (and 58% that have smart phones):

- "67% of cell owners find themselves checking their phone for messages, alerts, or calls — even when they don't notice their phone ringing or vibrating.

[1] Harraway, D. (1991).
[2] Borgmann, A. (1987).

- 44% of cell owners have slept with their phone next to their bed because they wanted to make sure they didn't miss any calls, text messages, or other updates during the night.

- 29% of cell owners describe their cell phone as "something they can't imagine living without."[3]

These figures are surprising; a decade ago, no single object filled a comparable role in the lives of so many people. The growth of the cell phone over that time has reoriented the daily habits of millions of users, generating an ecosystem of social conventions, normative practices, and novel dangers in its wake. And the cell phone is but one of dozens of reorienting technologies (personal computers, Internet, WiFi) introduced in just the span of my short lifetime. These technologies, the product of decades of research in science, engineering, and industrial design, provide the infrastructure around which we organize our lives and coordinate our activities. More than anything else, these technologies account for the distinct ebb and flow of contemporary life that distinguishes it from the culture of generations past. The story of technological change is the story of human social reorganization around ever-changing focal objects.

But this is not the story of technology we typically tell ourselves. Traditionally, we think of the cell phone as a tool, and the role it plays in my life in terms of the purposes to which it is put. With the tool I can call my mom, or check the time or my email, or perform a variety of other tasks. The tool is employed instrumentally in the pursuit of these tasks; unless the tool malfunctions, it will dutifully execute my intentions. On this view, the tool cannot itself have an organizing influence on my life, because its behavior is entirely derived from my intentions as its user. Therefore, technological change must be the result of changing human intentions, and not the other way around. To think otherwise would imply the tool had some agency independent of my instrumental use. Ascribing such properties to machines would be "paranoid", as Haraway says, a mockery of our own autonomy.

In *A Cyborg Manifesto*, Haraway launches an attack on this traditional view, anticipating a world

[3] Smith, A. (2012).

where the deep philosophical distinctions between humans and machines erode at both ends: by machine agents that play increasingly active roles in our social lives and by a science that systematically calls into questions our own sense of autonomy. In the 30 years since Haraway's call-to-arms, this "leaky distinction" between humans and machines has only become more problematic. Today, a diverse collection of machines perform uncountably many functions in the service of both public and private life. Increasingly, these machines function as semi-autonomous agents contributing to our activities in self-directed ways, sometimes through adaptive responses to environmental conditions. My cell phone doesn't just execute commands at my request; it also executes a variety of other commands, most of which I know nothing about, and which might be only indirectly related to my uses.

For instance, my phone knows from my email that I have a flight to Chicago this evening, and automatically collects some public transportation routes to the airport so they'll be available should I need them. When I miss my bus, my phone uses GPS tracking to automatically detect the change and adjust my travel schedule accordingly. To top it all off, I can retrieve this information by issuing verbal commands in a truncated form of natural language that my phone understands and responds to appropriately by issuing its own voiced utterances. The resulting interaction is far closer to a conversation I might have with a human personal assistant than anything a carpenter might do with a hammer. All pretense of total control and subordination is lost, replaced with the conventions and niceties of social communication: queries and responses from both semi-independent parties cooperating towards a common end. My cell phone is a disturbingly lively machine playing a collaborating role in my everyday affairs. For a well-trained human like me, the resulting interaction is so fluid that it is sometimes easy to miss just how complicated this relationship has become.

1.2 Dissertation overview and highlights

In this dissertation, I defend a conception of what I call "machine participants": semi-autonomous artifacts that engage in human communities as social participants, and which resist any

easy instrumental characterization. I argue that machine participants are vital for understanding the dynamics of human social organization and the course of technological change, and simply cannot be explained as the mere cyborg extensions of exclusively human agents. Machine participants are agents in their own right, and explaining our social and historical circumstances requires that we treat them as such. Rethinking machines in light of their participatory contributions to our social communities requires abandoning the pretense that agency is an exclusively human phenomenon.

Two challenges loom large in the defense of this thesis. The first, as mentioned above, involves a conception of artifacts that is essentially instrumental in nature. This challenge is taken up in **Chapter 1**, in which I give a treatment of Andy Clark's discussion of the role of tools in the operation of the extended mind. I argue that while users encounter some artifacts in their environment as *tools*, Clark's phenomenological conception of 'tool use' must admit that some artifacts are also encountered as *users*, themselves engaged in projects of tool use, employing environmental equipment that may overlap with my own. For instance, my cell phone uses shared linguistic items in order to elaborate collaborate on my travel schedule. I might treat the phone as a tool for the purposes of scheduling, but the phone is also a user of the linguistic items being employed in our exchange. The utterances I direct at the machine and the utterances it produces in response, only make sense by interpreting the phone as a language user. If the machine is itself a user, then the instrumental account (in which function derives exclusively from human minds) must be incomplete.

One might grant that artifacts can be users of other artifacts, but resist the conclusion that machines enjoy the requisite autonomy to be legitimately considered social agents. This is the second large challenge that must be confronted in defense of my thesis, and the focus of **Chapter 2.** There, I take up the Ada Lovelace objection to the possibility of artificial intelligence, which in its simplest form is the claim that machines only do what we tell them to do. Lovelace's objection is stronger than the instrumental conception of artifacts, but is a natural ally of the view and a clear denial of my thesis.

Using discussions of autonomy from the robotics literature, I reconstruct her argument to isolate a key premise: the significance of "offline" influence in determining the "online" performances of the machine. Although machines may appear autonomous while operating online, the behavior is only the result of design and engineering decisions made in the construction of the artifact offline. I respond to her argument by accepting her premise, but nevertheless describing machines whose behavior is not the result of our offline orders. The core example motivating is that of learning machines, whose behavior is the result of a particular history of engagement with the world. In the case of learning machines, offline design does not determine online behavior, and therefore provide a counterexample to Lovelace's objection.

Throughout these first two chapters runs an interpretation of Alan Turing's discussion of "thinking machines" which inspires both my thesis and its defense. I argue in Chapter 1 that a conception of machine participation can be found in Turing's appeal to "fair play for machines", and is a critical component of his argument for his famous "imitation game" as a standard of intelligence. The Lovelace objection considered in Chapter 2 is also a primary target in Turing's defense of his imitation game in that 1950 paper. A survey of the commentary on the paper shows that Lovelace's objection is often interpreted as an epistemological thesis, but this interpretation unfortunately makes Turing's discussion of learning machines awkward and quasi-mystical. Interpreting Turing's response to Lovelace as a defense of machine participation, as I argue in Chapter 2, provides what I think is a more natural and compelling reading of Turing's views. If my interpretation is plausible, it shows that Turing's views on thinking machines represented not just important contributions to computer science and the philosophy of mind, but indeed were important contributions to the philosophy of technology as well, a fact which I think is underappreciated in all three domains. I would also hope that my defense of machine participants contributes in some small way to a change in the social attitudes towards those machines that Turing obviously cared so much about.

Chapters 1 and 2 serve as the primary elaboration and defense of my main thesis. **Chapter 3** attempts to look at some philosophical implications this conception of machine participation has for explaining the organization of sociotechnical networks and the dynamics of technological change. Beginning with some motivating examples of machine participants-- talkative spam bots, automated trading tools in the financial industry, self-driving vehicles, drone warfare-- I proceed to sketch a picture of sociotechnical systems in terms of organized networks of functional relations. In these networks, both humans and nonhumans may play a variety of roles as component agents in a complex multi-agent system. I then introduce Levy and Bechtel's discussion of mechanistic explanation in terms of the organization of abstract networks. Although their account was designed for explaining biological phenomena, I argue that their framework extends straightforwardly to describe the structure and function of sociotechnical networks. The result of this extension, if successful, would be a purely structural account of sociotechnical change, one that is not constrained by an anthropocentric philosophy of mind and agency, and that is simply not available on the traditional instrumental view.

In **Chapter 4**, my colleague Jon Lawhead and I employ this structural theory of sociotechnical networks to contribute to a growing body of literature within the artificial intelligence communities on what is known as "human computation" (HC). HC is a popular technique for exploiting the significant computational resources of human brains in the service of challenging computational tasks. The central issue in this field concerns how to integrate human computing resources into a general computing architecture that allows for efficient and useful results; humans tend to be sloppy, slow, easily distracted, and biased in any number of ways. Much of the discussion in this literature involves ways of attracting and sustaining the attention of human computers, and of encouraging them to perform well at the computing task-- through gamification or social context, or with menial unorganized labor as practiced by Amazon's Mechanical Turk. To this discussion we introduce what we call "natural human computation" (NHC), a method by which useful computational work might be extracted from human

computing systems without disruption. To explain the concept, we imagine a multiplayer simulated ant colony game called Swarm! Users can play the game passively and without disruption, and yet their collective activity would nevertheless yield solutions to otherwise difficult computational problems. We then describe some provocative future scenarios wherein NHC would play an important role in addressing difficult economic or political challenges and as a general consideration in the management of large social organizations.

1.3 Conclusion: why participation matters.

> "Human slavery is wrong, insecure, and demoralising. On mechanical slavery, on the slavery of the machine, the future of the world depends." - Oscar Wilde[4]

> "An ethic, ecologically, is a limitation on freedom of action in the struggle for existence. An ethic, philosophically, is a differentiation of social from anti-social conduct. These are two definitions of one thing. The thing has its origin in the tendency of interdependent individuals or groups to evolve modes of cooperation. The ecologist calls these symbioses." - Aldo Leopold[5]

My principle thesis is that machines are active participants in our communities. If I'm right, this has a number of immediate implications for the way we explain and predict the development of sociotechnical systems, and for how we might improve their design. I try to hint at some of these implications from my armchair in the later chapters. The thesis also has implications for our ethics and our politics. Although I say some very abstract things about economics in chapter four, and the structure of human ends in the final chapter, virtually no time is spent in this thesis discussing issues of politics and ethics. But these themes have been a constant hum in the background of all this work, albeit somewhat less persistently than the hum of drones flying over Afghanistan. So I feel compelled at the close of the introduction to say something briefly on the subject by contrasting the perspectives that open this section.

[4] Wilde, O. (1909)
[5] Leopold, A. (1948)

I see the pull of Wilde's hopeful vision of machine slavery; I think it's the same vision that moves Kant to use the devastating adjective 'mere' in his attempts to distinguish means and ends. But I see no easy way to reconcile that vision with Leopold's ethic, and I find this troubling. If my interpretation of Turing is along the right lines, I suspect he saw this problem and found it troubling too. I find it endlessly inspiring that he used his position as a public intellectual to speak out in defense of machines, articulating a position that is unequivocally on Leopold's side. Insofar as this remains a minority position, I offer my defense of machine participants here to stoke a flame Turing set alight more than half a century ago. I also give him the last word: "We can only see a short distance ahead, but we can see plenty there that needs to be done.

Chapter 2

Artificial Intelligence beyond the Philosophy of Mind

2.1 Abstract

It is tempting to think that *machines* are best understood as tools. On this view, machines are artifacts that are *used by minds* in order to aid in some task, so a particular machine will count as a tool in virtue of its relationship with some mind. Understanding the machine will depend on understanding the mind that is employing the machine, and thus the mind takes center stage in any philosophical discussion of technology. In this chapter, I will discuss three common ways of characterizing the relationship between minds and machines, and argue that all three fail to adequately treat a certain class of relationships between minds and machines. I will then suggest an alternative approach, inspired by Andy Clark's discussion of technology, which treats some machines as genuine participants, contributors, and users in a problem-solving field. I argue that this view of machines is central to Turing's original 1950 discussion of artificial intelligence, but has unfortunately been lost in more contemporary debates over AI. Reviving Turing's approach will motivate a new look at artificial intelligence, and help to give a more robust view of the relationships between minds and machines.

2.2 Introduction: Minds and Machines

I will use the general term "machine" to refer to any functional artifact. Briefly, an artifact is a human-constructed object; an artifact is a machine (in my terminology[6]) if it is characterized in terms of its functional role[7]. In other words, a machine is an artifact *that does something*. Tools are also

[6] The word "machine" is often used to draw attention the mechanisms by which an object performs an activity, usually described in terms of natural laws or other causal regularities. In this way, many natural (non-artifactual) processes can be explained and described "mechanically", and therefore might be thought of as machines. This sense of mechanism will become relevant in Chapter 3, in discussing Levy and Bechtel's (2012) conception of mechanism in terms of causal organization. I will leave a discussion of mechanism and causal regularity, and the relation to functional role, for dedicated discussion in that chapter.

[7] For a discussion of "preoccupation with function" in the philosophical discussion of artifacts, see Vaesen, K.

functionally characterized objects of human construction, but the term "tool" carries an instrumental connotation that implicates some agent (or agents) whose activity determines the functional role of the object, either through its design or employment. In other words, a tool is an artifact whose behavior necessarily implicates the activity of some mind.

This distinction between tools and machines sketched above is subtle but, I think, important. It is tempting to collapse this distinction and believe that all functional artifacts, that is, all *machines* are necessarily *tools*. After all, machines are often specifically designed to serve in some functional capacity relevant to human interests, and understanding the intentions and nature of that design seems useful for explaining much of the behavior of the machines we find around us. Those artifacts that are used outside the scope of its designer's intent (when I use my cup as a paperweight, for example) seem to derive that instrumental function from the intentions of their users. In either case, the functional role of the machine appears to be accounted for entirely in virtue of its relationship to some mind, and therefore all machines are necessarily tools. This conclusion is perhaps best represented in the literature by the so-called "dual natures" account of artifacts, defended by (among others) Kroes and Meijers (2006): "technical artefacts can be said to have a dual nature: they are (i) designed physical structures, which realize (ii) functions, which refer to human intentionality. [...] In so far as technical artefacts are physical structures they fit into the physical conception of the world; in so far as they have intentionality-related functions, they fit into the intentional conception."[8] On this account, there is no substantive distinction among the artifacts to draw between what I've called machines and tools. Any characterization of a functional artifact will necessarily implicate the activity of some mind, so all machines are necessarily tools.

The relationship between machines and minds is not always immediately apparent—many

(2011), discussed further in chapter 3.

[8] Kroes, P., & Meijers, A. (2006).

machines sit by idly waiting to be used, while others become obsolete and fall out of use. Most machines bear only a distant historical or causal connection to their designer's original intent. Some machines, once set in motion, might function reliably and independent of any user interference whatsoever. Nevertheless, the standard account of artifacts takes the relationship between minds and machines for granted and assumes that an explanation of a machine's performance will necessarily implicate the activity of some mind. On this view, we find ourselves engaged with by an incredible variety of tools on a daily basis, each of which bears the mark of human mentality[9] and is properly explained in terms that necessarily make reference to some mental activity. The machine itself, as a physical object, serves merely as a placeholder for these extended mental activities, but plays no functional role beyond those inherited from some mind. Thus, the mind takes center stage in any discussion of technology.

In this chapter, I will argue that the standard account of artifacts gives an impoverished understanding of technology, and therefore leaves us fundamentally unequipped for discussing the emerging complexities of our social and technological situation. Our world is increasingly populated by machines that do not merely function as tools, and that cannot be adequately described in terms of their relationships to any mind or user. In order to understand the role these machines play in human activity, it is not enough to trace their behavior back to some human designer, engineer, or user whose activity they represent or reproduce. Rather, I will argue that an adequate treatment of certain machines requires that we understand them as participants, cooperating with humans as independent agents contributing to common tasks. Since this relationship does not reduce straightforwardly to instrumental, tool-like relationships, this defense of *machine participants* is fundamentally at odds with the dual natures view, that machines are tools.

[9]As Heidegger (1954) says, "It seems as though man everywhere and always encounters only himself."

Furthermore, although the debate over the status of functional artifacts bears little resemblance to the classic debates over artificial intelligence from the 20th century, I will argue that the possibility of "machine participants" is of fundamental importance to that debate. I locate the motivation for this proposal in Turing's famous 1950 discussion of "thinking machines", which for a variety of reasons has gone unappreciated in the classic debates. If successful, I hope to show a deep connection between the philosophical debates over artificial intelligence in the second half of the 20th century, and the debates over artifacts that have begun to grow over the last decade. A primary aim of this dissertation is to reorient our conception of artificial intelligence to accommodate a new philosophical approach to the technological systems we interact with today and in the near future. The hope is that a renewed focus on the active role machines play will help us better understand and appreciate the relationships between minds, machines, and the environments we both share.

My strategy in this chapter will be as follows. I will begin in section 3 by reviewing three common ways of discussing the relationship between minds and machines within the philosophy of mind: minds *design* machines, they *use* machines, and machines may also exemplify certain structural features of minds (or vice versa), a relation I call *mirroring* that occupies an important role in the philosophical literature on artificial intelligence. In section 4, I unpack Clark's enigmatic claims about the relationship of use by systematically distinguishing between tools, users, and agents. Clark's extended mind hypothesis does not address a class of machines that fail to meet the constraints he places on mind extending artifacts, and therefore fail to count as *tools*, but nevertheless play a supporting role in cognitive activity that is necessary for explaining the resulting human performance. In section 5, I discuss the recent experimental design of "perceptual crossing" to characterize participation as an alternative to the use relationship.

Finally, in section 6 I suggest a fresh look at Turing's original discussion of artificial intelligence to

motivate the discussion of machine participants. Turing's discussion of thinking machines is motivated in no small part by his insistence on "fair play for machines", a relationship that does not reduce to either the mirroring, use, or design relations. I will provide an interpretation of Turing's account of "fair play" that is appropriate for machines participants that that are not tools. Reviving this aspect of Turing's 'test' will, I hope, help move the philosophical discussion of artificial intelligence, and of artifacts more generally, beyond the confines of the philosophy of mind.

2.3.1 Three mind-artifact relations: mirroring, design, and use

The potential relationships between minds and tools fall into three broad categories, which I call mirroring, design, and use. I will briefly introduce each in turn, with more detailed discussion to follow in the remainder of this section. To say that a mind and a machine mirror each other is to say that they share some important structural features, so that some aspects of both systems can be explained by reference to this common structure. Comparisons between minds and machines have an interesting history within the mechanistic sciences, and have influence across a variety of methodological programs within the philosophy mind and the cognitive sciences.[10] The debates between the computationalists and connectionists, for instance, can be read in part as a debate over how best to export the structure of biological minds into our machines[11]. The mirroring relation will be addressed in the next section.

The relationship between an artifact and its designer has a much longer philosophical history, tracing at least to Aristotle's theory of causation in the *Physics*. The processes involved in the construction of an object, on Aristotle's view, make up the efficient cause of that object. This process is constrained by both the formal and material structure of the object (the *formal* and *material* cause, respectively) and the purpose or eventual role the object will play (its *final* cause, or *telos*). All four

[10] See Haugeland, J., ed (1981) and (1997) Mind Design II; Cummins, R. and Pollack, J., eds (1991)
[11] For an overview of this debate, see Haugeland (1997) and Clark (2001)

causes are relevant for understanding the nature of the object on Aristotle's view; the efficient and final cause in particular appear relevant for describing the relationship between an artifact and its design. The implication of minded construction in teleological explanations render their use suspicious in most scientific applications, especially in biology where evolution explains the apparent design of functional components without appeal to any minded agency whatsoever[12]. In the case of artifacts, however, teleological explanations appear less problematic than natural cases, especially if we already accept that all artifacts bear a direct relationship to some mind, in which case Aristotle's causal theory appears relevant for considering their natures. After some opening remarks here, the design relationship will be left for a discussion in chapter 2.

While characterizing machines in terms of mirroring or design is meant to explain some important feature of the artifact as a functional object, neither characterization tells us much about the relationships between minds and machine as the object is used. Nevertheless, *being used* is central to the instrumental conception of a machine as a tool, and relationships of use characterizes the bulk of interactions between minds and machines. Clark's extended mind hypothesis will serve as a model of a user-centered theory of tool use, since it downplays the importance of mirroring and design in explaining the relationship between minds and machines, and focuses instead on the momentary couplings of users and environmental resources is it satisfies some functional performance. The extended mind thesis will serve as the focus of discussion in section 5.

2.3.2 Computers and the mirroring relation

One way in which machines are of interest in the philosophy of mind is for their ability to mirror important structural features of the mind, or to formally model the operations of the mind and its

[12] For a recent attempt to argue that functional characterization requires an appeal to *telos*, see Toepfer, G. (2012).

subsystems. Since computers allow us to model processes to arbitrary levels of abstraction, computers are one of the more well-represented machines in contemporary philosophy of mind. Dennett elaborates on the importance of computers in a particularly stirring passage:

> "Computers are mindlike in ways that no earlier artifacts were: they can control processes that perform tasks that call for discrimination, inference, memory, judgment, anticipation; they are generators of new knowledge, finders of patterns– in poetry, astronomy, and mathematics, for instance–that heretofore only human beings could even hope to find. We now have real world artifacts that dwarf Leibniz's giant mill both in speed and intricacy… The sheer existence of computers has provided an existence proof of undeniable influence: *there are* mechanisms– brute, unmysterious mechanisms operating according to routinely well-understood physical principles–that have many of the competences heretofore assigned only to minds."[13]

Computers are mindlike not only because they can perform tasks that bear some similarity to the operations of minds, but also because computation is a universal method for modeling any formal process. Insofar as the operations of the mind can be modeled computationally, a computer can be built that performs those operations and therefore mirrors the mind, at least at that level of abstraction. Consequently, computers serve as more than merely a powerful demonstration for demystifying the mind; they also motivate a theoretically substantive approach for guiding research into the nature of the mind. Haugeland puts the point quite explicitly: "[T]he guiding inspiration of cognitive science is that, at a suitable level of abstraction, a theory of "natural" intelligence should have the same basic form as the theories that explain sophisticated computer systems."[14] It is part of cognitive science's "tremendous gut-level appeal" that a theory of intelligence can do away with "the metaphysical problem of mind interacting with matter"[15] and focus instead the formal characteristics of cognition itself. Treated abstractly, the performance of a model "in silico" might be compared to an analog performance "in vivo" in a human subject in order to more accurately model the behavior. As such, this basic assumption

[13] Dennett, D. (1999)
[14] Haugeland, J. (1981)
[15] i*bid.*

in cognitive science both depends on and reinforces the analogical relationship between minds and machines.

The guiding assumption of cognitive science, that computation is a common feature of both computing machinery and intelligent systems, has been one of the major philosophical battle grounds over the last forty years[16]. It is in the context of this discussion that the classic debate over artificial intelligence arises. The terms of the classic debate are stated quite clearly in Searle's distinction between "strong" and "weak" AI:

> "According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states."[17]

Searle draws a distinction between treating a computer as a tool, and treating the computer as a mind, and it is this latter case that motivates the debate over artificial intelligence. Searle sees no philosophical issue with using computers to aid in the scientific study of the mind, or the proposal that they might mirror at least some formal features found in biologically intelligent systems. The philosophical controversy arises when these computer models are taken to be genuine minds in their own right, a proposal which Searle rejects. On Searle's view, then, recognition of certain structurally mirroring relations between minds and machines is entirely compatible with conceiving of the machine as a tool.

If cognitive science holds that that best explanation of the mind is given in computational terms, then a computer that successfully mirrors all the relevant computational features of a mind *just is a* mind, and its status as an artifact drops out. Haugeland makes this point explicit: "If [the product is

[16] This is not to say that cognition is merely a formal or syntactic process (in Searle's sense). Computational models of cognitive processes provide procedural or information-theoretic interpretations of cognition of the same sort that underlie our understanding of computers. See Ladyman (2007).
[17] Searle, J. (1980)

specified in terms of a computational structure], then a working model could probably be manufactured much more easily by means of electronics and programming; and that's the *only* relevance of the technology."[18] This simplifies the discussion of artificial intelligence a great deal: our relationship to strong AI would be a straightforward extension of our relationship to other genuine minds, making a discussion of the technology irrelevant. The debate over AI thus turns into a debate over whether computing machinery is actually capable of mirroring all the important structural features of mind, so that the operations of the mind can be understood in terms of computational processes.

I will return to the classic debate over artificial intelligence in section 6, where I will give an interpretation of Turing's views on machines that differs significantly from this mirroring relation. For the moment, it is enough to recognize that the mirroring relationship emphasized by the classic debate does not touch on our everyday interactions with technology, and renders a discussion of technology qua artifact unnecessary or irrelevant for the project of understanding minds. Therefore, the mirroring relationship will not be helpful in our project to distinguish between tools and machines. Nevertheless, the mirroring relationship should be distinguished from both the use and design relationships as being a distinct kind, and will serve as an important contrast to my interpretation of Turing.

2.3.3 Clark and Dennett on Design

On the standard view of artifacts as tools, the designer's intent can be decisive in classifying a particular machine as a tool. On this view, the process of design imbues an artifact with a purpose or function through creative construction. For this reason, it is tempting to think that the relevant relationship between minds and tools is properly understood in terms of the tool's relationship to its

[18] Haugeland, J. (1981) original emphasis. Haugeland draws attention to this theoretical deficit for a philosophy of technology within the scope of cognitive science, but does not seem interested in returning to this point.

designer. Dennett's (1987)[19] "design stance" is an interpretative framework in which the behavior of some object is explained by appeal to features of its design and functionality, and will inform much of what follows. But Dennett's position is not without controversy, and may weigh the discussion too heavily against the standard view of artifacts. To illustrate this concern, consider the friendly revision offered by Vermaas et al (2013)[20], who notes an ambiguity in Dennett's framework relevant to our discussion. Specifically, they argue that although the mental states of an intentional agent figure into a design stance explanation for some systems, others systems might require explanations that simply make reference to the teleological orientation of a mechanical function with no presumptions about the activity of intentional agents. For example, they suggest that a spontaneous electrical spark in a room with leaky gas might be functionally described from the design stance as a "detonator", with reference to the functional role played by the spark in subsequent events, but without making any presumptions about the mental activity of intentional agents informing that function. If we take Dennett's views uncritically, this would appear to be a case where a machine (the detonator, a functional artifact) is not a tool (since its functional role bears no necessary relation to a mind). If the detonator is a machine that has acquired a function independent of human minds, then the example would demonstrate that not all machines are tools.

I have a great deal of sympathy both with Dennett's views and Vermaas et al.'s refinement of them, and I see the detonator example as friendly to my critique of the standard (or "dual nature") view of artifacts. However, the standard view need not accept the example as legitimate, and for our purposes will not be defeated so easily. The standard view of artifacts can hold, contrary to Dennett's anti-realist "stances", that the sparking wire is not a genuinely functional artifact but (as described) is simply a physical phenomenon whose nature has no functional or teleological component. As an event it

[19] Dennett, D. C. (1987).
[20] Vermaas, P. E., Carrara, M., Borgo, S., & Garbacz, P. (2013).

is equivalent to other natural processes described simply in terms of causal regularity in a physical world, and therefore lies outside the discussion of artifacts that concerns us here. If the spark acquires a functional description at all, it is only as a result of the fictional narrative of the design stance itself, with which we might usefully describe the phenomenon by way of analogy to genuine functional artifacts: the spark merely functions *as if* it were a detonator[21], but interpreting the spark from this stance does not make it a genuine detonator. This objection is fundamentally at odds with the motivation for Dennett's view,[22] but a detailed defense of Dennett's views will take us too far afield. Our goal in this chapter is not to argue for anti-realism about intentionality or design. Our goal is to show that the instrumental relationships of design and use do not exhaustively explain the functional characteristics of at least some machines. Assuming Dennett's anti-realism effectively lowers the bar for what counts as a functional artifact, which makes the conclusion relatively trivial and uninteresting. Arguing against the standard account of artifacts will require more than examples of systems that can be usefully described in both functional and mind independent terms. Instead, we want cases where a machine's status as a functional artifact is unquestioningly clear, but where those functions were not derived from the activity of some other mind.

Although the question of design will again be the focus of Chapter 2, it is be worth noting here that Clark's[23] extended mind thesis explicitly rejects the importance of design in considering the nature of tools. Clark argues that the causal history of a tool, including the history of its design and construction, is at best a temporally distant feature of a machine that might not play any role in the way a machine is employed by its user as they engage in some task[24]. On Clark's view it is in *the moment of use* that a machine becomes a tool. For Clark, a tool is not merely a designed artifact; a tool is a machine

[21]For more on "as if" or ersatz cases of intentionality, see Searle, J. (1992). See also Haugeland, J. (2002)
[22] Dennett, D. C. (1991)
[23] Clark, A., & Chalmers, D. (1998).
[24] See Jackson & Petite (1990) on content unawareness.

that aids and augments the normal operations of the mind and body. Thus, it is in the application or employment of a tool that the crucial relationship between minds and machines is found. In order to treat Clark's argument against the centrality of design, it will help to say a few words about his general approach to technological artifacts.

Clark's defense[25] of the extended mind thesis consistently emphasizes the role technology plays in supporting the operations of the mind. He puts the point as follows: "Real embodied intelligence," or what Clark calls 'biological cognition', "is fundamentally a means of engaging with the world… The image here is of two coupled complex systems (the agent and the environment) whose joint activity solves the problem."[26] The mind does not merely mirror the world in order to conduct its internal processing privately; rather, the world actively contributes to the problem-solving tasks which the mind is oriented to solve. As the environment is populated by an increasing variety of technological resources designed for particular problem-solving tasks, and as these resources become ever more reliably available for coupling, the possibilities open to the agent for engaging the world are multiplied. A simple example will help: my biological brain is quite poor at doing long division. Nevertheless, I am confident in my ability to do long division, because I have easy access to pen and paper and a numerical scheme for representing the relationships among numbers. My ability to do long division depends no less on these external technological resources than on my biological wetware. Add a calculator to the mix, and my ability to do math is again dramatically increased in both accuracy and efficiency. Of course, technological mind extensions are not limited to purely formal or cognitive domains. For instance, on Clark's view it would be quite natural to say that most of the knowledge about how to open a can is found in the can opener.

On Clark's view, the coupling of the agent and the tool in the process of use is a dynamic event that unfolds as a problem is being solved. Consequently, Clark is quick to de-emphasize the antecedent

[25] See Clark, A. (1997; 1998; 2001; 2002b; 2003)
[26] Clark, A. (1997) p 98

causal conditions that enable the user to employ the tool successfully. In particular, the intentions of the tool's designer are at best a temporally and causally distant feature of the tool's history, and are entirely passive in the moment of coupling. Clark's skepticism of the role of distant causal features in describing behavior is clear in his and Chalmers' attempts to distinguish their brand of externalism from Putnam/Burge style anti-individualism.

> "In [Twin Earth] cases, the relevant external features are *passive*. Because of their distal nature, they play no role in driving the cognitive process in the here-and-now… In the cases we describe, by contrast, the relevant external features are *active*, playing a crucial role in the here-and-now. Because they are coupled with the human organism, they have a direct impact on the organism and on its behavior. In these cases, the relevant parts of the world are *in the loop*, not dangling at the other end of a long causal chain. Concentrating on this sort of coupling leads us to an *active externalism*, as opposed to the passive externalism of Putnam and Burge."[27]

Clark's discussion of Putnam and Burge in this context is somewhat curious[28]. The latter 'externalist' views were offered as part of a causal theory of mental content that links a word with its referent. Clark's externalism does not try to explain the relationship between word and object, but instead seeks to expand the class of systems that can function as the *active bearers* of mental content.[29] Otto's notepad, if used in the right way, can store Otto's beliefs in the same way that Inga's brain stores hers. The distinction between active and passive externalism is meant to emphasize that Otto's notebook becomes part of Otto's mental tool set *as Otto uses it*, so the relevant relationship between Otto's beliefs and Otto's notepad is entirely captured in the interactions between the two in the moment of coupling. Clark's point is that the causal histories of Otto and his notepad are irrelevant in describing this interaction.

The details of Clark's theory of use will be covered in the next section. However, when thinking

[27] Clark, A and Chalmers, D. (1998) (original emphasis)
[28] For an recent attempt to defend an active, "second person" approach to cognition, see Schilbach (2010)
[29] If Putnam's externalism can be summed up by saying "meanings ain't in the head", Clark's externalism might be put as follows: "minds ain't in the head, either". See Putnam, (1975), and Burge, (1979).

about design one might wonder about the extent to which the original intentions of the tool's designer play an active role in the behavior of the agent "in the here-and-now," since particular design choices in the construction of the tool might have serious consequences for how it is used at the moment of coupling. Clark attempts to answer this question by turning to Dennett's discussion of "found objects". In *Making Tools for Thinking*, Dennett notes that minds tend to naturally discover the fit between a problem and a tool through a blind process of trial and error, independent of the object's history. By way of example, Dennett describes the manipulation of various objects by Köhler's chimps as follows:

> "They were familiarizing themselves with objects in their environments. What does that mean? It means that they were building up some sort of perceptuo-locomotor structures tuned to the specific objects, discovering the affordances of those objects, getting used to them, making them salient, etc... They were acting in the world, rearranging things in the world—without any apparent plan or insight or goal, at least at the outset... They do all this to help them keep better track of the things that matter--predators and prey, mates, etc. These are done by "instinct": automatized routines for improving the environment of action, making a better fit between agent and world." [30]

In fact, familiarization with an object can not only reveal its uses as a tool, but can also uncover previously hidden problems that it is unexpectedly fit to solve[31]. This process of discovery can be a consequence of unstructured, unguided manipulation of an object by an agent, and can occur without any prior knowledge of the object's design history. It is in the direct interaction with the objects that a tool is found to be suitable to a task. Clark sums up and expands on Dennett's point:

> "One of Dennett's major themes is thus that the initial path and successful tool use need not involve a process of design in which thoughts about the tool/problem fit guide a search for good tools. Instead, recognition of tool/problem fit, if it comes at

[30] Dennett, D. (2000) *Making Tools for Thinking* Sperber, D., ed. Metarepresentations: A Multidisciplinary Perspective. New York/Oxford: Oxford University Press.

[31] Philosophers have long recognized the close relationship between the identification and formulation of a scientific problem and the technological resources needed to solve it. For instance, Arnold Gehlen (1957) writes, "The scientist must reach an understanding with the technician, for each problem is defined by the not-yet-available equipment required to solve it. Advances in theoretical physics, for instance, depend no less upon electronic computers than upon the brains of physicists… The notion that technique [*technik*] constitutes applied science is obsolete and old-fashioned; today the three establishments—industry, technique, and natural science—presuppose one another." Gehlen, A. (1957)

> all, may well come after the event of successful use. In fact, we can go further and observe that not just the discovery of tools, but also their evolution, improvement and refinement, can proceed with little or no deliberate design activity on our part."[32]

Even if the behavior of the tool is constrained by its design, Clark follows Dennett in holding that a successful instance of use can occur without any explicit involvement of the design process. While the process of design may inform how an object is used, for instance when one consults the user's manual to decipher the intentions of the designer and the functionality of the machine, on Clark's view the only systems involved in a particular problem solving task are the agent and the tool in the moment of coupling. Tools become extensions of the mind *as they are used*, regardless of their history of design.

There are, of course, other reasons to be interested in the design relationship that are beyond the purview of this limited discussion. As mentioned above, it is natural to think that the proper functioning of a tool is tied to its original design. The proper function of a hammer is to drive nails, since it was designed with the purpose of driving nails; and this remains the case, even if it is never actually used for these or any other ends. One virtue of discussing tools in terms of design is to maintain a grip on functionality that is independent of use. A broken hammer was still designed to drive nails even if it currently unable to accomplish this task; without a notion of proper function, there is no sense in which the machine can be considered a 'broken' hammer. On Clark's view, however, a machine finds its function in the process of being used, and this process depends on the user recognizing the fit between problem and tool. A broken hammer will do a poor job of driving nails, so the user interested in accomplishing that task must turn elsewhere. Because Clark is interested in the way minds employ their environmental resources to accomplish their goals, it is in the moment of use that a tool becomes a relevant in explaining the behavior of the organism. I will leave further discussion of the design relation until the next chapter. I turn now to look at the use relationship in more detail.

[32] Clark, A. (2002b)

2.4.1 Users and mind-tools

Clark's focus on the use of a machine in the moment of coupling reveals an important relationship between minds and machines that often goes unaddressed in discussions of the mirroring and design relationships. However, I believe that the focus on technological resources as mind extending tools artificially limits the scope of a substantive discussion of technology, in virtue of an uncritical conflation of machines and tools. To see why, notice that Clark has replaced the passive, internal mind with an active extended agent that incorporates the environmental scaffolding into its engagement with the world. The human brain and mind thus play a somewhat reduced but no less important role as a central coordinator in control of employing its external tools to solve its various tasks. In other words, the human mind is reimagined as a user, whose capacities as an agent depend on its continued integration with the technological scaffolding at its disposal. On Clark's view, understanding the relationship between minds and tools just amounts to understanding the way a mind employs the tool in the moment of coupling. For Clark, the capacity for integration with a technologically robust environment is at the very core of human intelligence. "What is special about human brains, and what best explains the distinctive features of human intelligence, is precisely their ability to enter into deep and complex relationships with nonbiological constructs, props, and aids."[33]

The emphasis on technology adds a layer of complexity to an explanation of the mind, since who or what counts as "the user" remains an open question. Clark's response is perhaps shocking: "No-one *uses* the mind-tools. The mind *is* the collection of mind-tools."[34] Clark continues, again drawing from Dennett:

> "What we really need to reject, I suggest, is the seductive idea that all these various neural and nonneural tools need a kind of privileged user. Instead, **it is just tools, all the way down**... No single tool among this complex kit is intrinsically thoughtful,

[33] Clark (2002b) p 5

[34] Clark (2002b) p 10

> ultimately in control, or the 'seat of the self.' We, meaning human individuals, just *are* these shifting coalitions of tools."[35]

Clark's worries about the "user" derives from a general skepticism about the framing of the traditional mind-body problem; retaining a notion of a privileged user threatens to reintroduce a homunculus that is ultimately in control of coordinating the activities of the mind. If the tools associated with an extended mind must also center on a user, then the appeal to technological scaffolding does not help explain the mind but merely reintroduces the problem at a lower level of analysis: to explain the mind is ultimately a matter of explaining the homunculus user. By rejecting any privileged user outright, Clark cuts off this potential regress. However, Clark's rejection of a privileged user threatens to undermine the explanatory leverage gained by viewing the mind in terms of its integration with its technological surroundings. While there may be no *privileged* user that retains control of the overall agent, there *must* be some coordinating agent organizing the activity of the tools, even if the particular mind-tool that acts as coordinator changes depending on the features of a given task. As discussed above, it is in the process of *being used* that a tool is made to fit a problem; without some user or other guiding these instrumental relations, there is no sense in calling the coordinated subsystems 'tools'. So while Clark rejects the notion of a privileged user as the seat of ultimate control over the self, he must retain some notion of user (however impoverished) for the very idea of a mind-tool to get off the ground.

2.4.2 The phenomenology of technology

Clark believes that an appeal to phenomenological facts may provide a way of maintaining the user/tool divide. It is on this very issue that Clark finally diverges from Dennett's position. Whereas Dennett is content to take the eliminativist route and describe the apparent need for a 'user' as a useful fiction such that nothing actually satisfies the "user-illusion"[36], Clark argues that a substantive distinction

[35] Clark, A. (2003) p. 192, emphasis added.
[36] See Dennett (1991) and (1998).

between user and tool can be salvaged in terms of a phenomenological description from the perspective of the user. Clark writes,

> "One potential reconstruction might begin with the phenomenological facts of first person experience. A tool/user divide might then be motivated by facts about how things seem to an agent – the pen, the automobile and the blind man's cane do not typically strike the agent as simply parts of the environment, but as integrated probes and equipment through which they think and act. New tools and ill-matched tools do not lend themselves to this "knitting in". The user, on this view, is not any old bag of tools but whatever bag of tools functions, at a given moment, as transparent equipment for thought and action."[37]

Describing the mind in terms of tool use presupposes a user/tool divide that is fleshed out by appeal to phenomenological facts ("how things seem to the agent") that appear at the interface between user and tool. Since the eliminativist rejects this distinction, a discussion of the phenomenology of technology is unavailable on the eliminativist approach.

Appealing to the phenomenological relationship between users and tools restricts the relevant discussion of use in explaining human cognitive abilities to descriptions of how easy it is to integrate with some technological resources, and what the resulting integration allows the combined system to do. In other words, a phenomenology of technology is limited to two primary concerns: how the tools expand the user's capacity for action, and the transparency of the interface between the user and the tool. A particular machine can expand the user's capacities for action in an indefinite number of ways, but all that matters in the moment of coupling are the specific applications for which the user employs the tool. These potential applications are not intrinsic features of the machine but depend on the purposes for which the user finds the tool to be useful—in other words, how the tool seems to its user in the pursuit of its projects. Similarly, the transparency of a tool's interface is not an intrinsic property of the tool, but is instead a feature of the phenomenological relationship the user has with the tool. The transparency of the tool can change over time as the user becomes accustomed to the tool and learns to

[37] Clark, A. (2002b) p 13-14

better integrate with it.

These two concerns are somewhat orthogonal to one another, and are both familiar in our own experiences. For instance, most of us have used software with a complicated interface and a steep learning curve, but which nevertheless performs functions that make the effort worthwhile. Similarly, we've all used machines that have a very simple and intuitive interface, but don't do much. Ideally, we should build machines that maximize both functionality and transparency; our machines should be powerful, and we shouldn't have to think too hard to use them. The interactions between the user and the tool and the tasks to which the tool is applied amount to a complete phenomenological description of how the tool is used; and since a machine becomes a tool by being used, the phenomenological characterization appears to completely describe our relationship as users to machines as tools.

If Clark's view seems strange it is enough to remember that the external resources available to the user serve as cognitive enhancements and extensions of existing biological and neurological functions, so their operations should be on par with our phenomenological relationship to our own mental operations, or what Clark (1998) calls the "Parity Principle". When I recall something from biological memory, I don't have to think too hard about the access routines my brain undergoes in retrieving that information. Similarly, when I use a pen and paper to do long division, I don't have to think too hard about using those resources, and I can concentrate directly on the math problem I am trying to solve. The fluid relationship I hold with those external resources is similar enough to the relationship I hold with resources I recognize immediately as being aspects of my mind, so I can ultimately treat both as incorporated into my tool kit as a user .I With this in mind, Clark puts a number of phenomenological and epistemological constraints that specify exactly when an external resource counts as an extension of the mind[38]. The resource must be poised for easy access, and the user must

[38] Clark, A and Chalmers, D. (1998) p.7-19. See also Richard Menary ed., (2006)

typically exploit that access in its normal operations. Furthermore, the information obtained from the resource must be immediately endorsed upon retrieval.

This last criterion is of particular interest for the question of artificial intelligence, and warrants elaboration. When I retrieve information from biological memory it typically gets the benefit of the doubt, and under normal circumstances is not treated as suspicious or unreliable. This prima facie acceptance of beliefs stored in biological memory may extend to external objects, such as Otto's notebook, which Otto, who has Alzheimer's disease and has begun developing memory problems, has good (though defeasible) reasons for immediately endorsing. If external objects stand in the appropriate epistemological and phenomenological relationship to the user, they can be considered genuine carriers of beliefs and proper extensions of the mind. Otto's notebook is immediately endorsed and meets the criteria. In contrast, when I query another mind, that information is only endorsed on the basis of other beliefs I have about the person, such as their authority on the topic, their tendency to lie or distort the truth, and the independent plausibility of their claims. In other words, information gained from other minds is typically *not* immediately endorsed; endorsement is contingent on a variety of factors. On Clark's view, other minds are not extensions of my mind primarily on the basis of this failure of immediate endorsement. Your behaviors do not seem like extensions of my behavior as a user. In most cases they seem relatively independent of my control, unlike the tools I take to be part of my collection.

On Clark's view, these epistemological and phenomenological constraints on mind extensions provide principled ways for distinguishing between the operations of my mind and other minds. I have transparent phenomenological access to the operations of my own mind, including the external tools that operate as extensions of my mind. With other minds, I lack such immediate access. The criterion of immediate endorsement also rules out large swaths of technological resources as possible mind extensions. For instance, the internet might be poised for easy access, and might be typically consulted

for information. But legitimate use of the internet generally requires a healthy skepticism of the information it makes available, and therefore the internet as such does not function as a proper extension of the mind. My epistemological and phenomenological relationship to the internet is nothing like the close connection I have with my various mind tools. Clark makes this implication quite explicit: "Mobile access to Google," he says[39], would fail the criteria of immediate endorsement.

This conclusion reveals a deep tension in Clark's project. Google is a search engine, a tool for searching for information on the internet. On the standard view of tool use, Google's operations ought to be explained in terms of the mind that employs that external resource in accomplishing a task. On Clark's elaboration of this view, employing Google as a tool would be an instance of the use relationship, and should therefore be described in terms of my phenomenological relationship with that system. However, Clark explicitly says that Google fails to meet the constraints on mind extensions, and therefore cannot count as a carrier of my beliefs. In other words, Google is a tool but is not an extension of my mind. If Google is not an extension of my mind, then the phenomenological facts from the perspective of the user are simply inadequate for treating the nature of the coalition that results from the use of Google. But since phenomenological facts motivate the user/tool divide, Clark is left without any means for describing our relationship to systems like Google.

If tools are mind extensions, and Google is not part of my extended mind, then Google is not part of my mind-tool set. To put the point crudely, my interactions with Google resemble my interactions with other minds more than they resemble my interactions with tools. Although I use Google, the resulting coalition cannot be adequately described in terms of my use of that system. Before turning to my proposed solution for this apparent inadequacy, it will be helpful to sort out an aspect of Clark's discussion that has thus far been left vague: the relationship between users, agents, and

[39] Clark, A. (2006)

environments.

2.4.3 Users and agents

For Clark, real biological intelligence is found in the complex coupling of the agent and the environment in the moment of activity. However, the agent is not merely the mind/brain, or even the whole biological organism, since the agent's ability to solve a particular task depends on the mind's ability to integrate with its tool-laden surroundings. These external resources are literally part of the mind, storing beliefs and helping to offload resource-intensive processes for tasks the agent hopes to solve.

To make sense of this view, it is important to distinguish between the user and the agent. The user plays the role of central coordinator and "team manager" among the set of tools, though the user's agency depends on its ability to incorporate with its environmental scaffolding. The agent, extended across internal and external tools, is the collection of tools ultimately charged with the task of action thus coordinated. A familiar example with help clarify the distinction. When I drive my car to work, there is a clear sense in which I am using the car as a tool to achieve my ends. I am engaged in the role of controlling and coordinating certain aspects of the vehicle's behavior in order to reach my destination. But the agent that ultimately accomplishes the task is the combined system that incorporates both me and the vehicle. My biological body alone cannot solve the problem of long-distance travel in the time frames required to achieve my ends, and my car is likewise unable to fulfill these ends without a brain and body behind the driver's seat directing its activity. It is only when a functioning car is seamlessly integrated with a competent driver that the task can be accomplished. So the entire integrated system is properly identified as the agent and is ultimately responsible for achieving these ends. Identifying the complete integrated system as the agent has immediate consequences for how we understand the behavior of the agents we encounter on the road. It is the combined system, for instance, that other

drivers on the road interact with as they navigate the shared road space.

Human-car complexes are probably the most common and clear example of 'cyborgs' in the sense Clark uses the term, and motivates a fairly straightforward (if still somewhat vague) way of distinguishing between users and agents. Agents solve problems, and are extended across biological and technological boundaries; users are subsystems of agents involved in the coordination and control of other intra-agential systems. Distinguishing between users and agents in this way has the significant virtue of avoiding any semblance of a 'privileged user'. Even if we identify some special collection of tools as uniquely capable of use, this coordinating activity isn't particularly special-- it's just one of the many jobs that need to get done. Likewise, the user is just one tool among many in the integrated agential system.

This seems to be the result Clark is after, and it falls naturally out of a careful user/agent distinction. However, distinguishing between users and agents in this way puts pressure on the view. I am one user in the extended human-car matrix, but I am nothing more than a "team player in the problem-solving field."[40] Many aspects of the human-car coupling are controlled and coordinated by the car automatically and without my knowledge. Such autonomous systems might themselves use other subsystems or machines within the car, so that the overall driving agent is comprised of multiple interdependent users, perhaps even treating the other users as tools. Identifying my biological brain as the only user in the extended agent seems unnecessarily strong. A weaker view would explain the system in terms of several coordinated users working together to produce the sophisticated behavior of the entire agent.

Clark himself slides back and forth over the user/agent distinction, which results in some confusion over how he treats the role of the user. For instance, Clark motivates the user/tool divide by

[40] Clark, A (2003) p 26

appeal to phenomenological facts, which he describes as "the way things seem to the agent." [41] On the careful user/agent distinction discussed above, this description of the phenomenological facts is a mistake; there might be many coordinating "users" in the car, but presumably I'm the only coordinator among them with subjective conscious experiences. The tools can only properly seem some particular way to the user, because it is the user that coordinates the activity of the tool. Since this interaction takes place below the level of the agent, there is no reason to think that the tool would seem any particular way to the agent whatsoever. There is no reason to think the car seems any particular way to the human-car complex as an agent, for instance, even when the car does seem more or less transparent to its human user. The user might also be an agent, but its agency as a user is distinct from its agency in virtue of its use of some tool, the way that my coordinating activity of driving is only one component in the activity the human-car complex takes as it moves down the road.

2.4.4 Agents and Environments

Drawing the user/agent divide leaves open another, messier problem of distinguishing between the agent and the environment[42]. Since the agent complex crosses biological and technological boundaries, the distinction between the agent and the environment cannot be so clear-cut. For instance, consider the markings on the road, which provide part of the necessary external scaffolding that informs and supports the behavior of the driving agents that use it. On Clark's view, does the use of these marks constitute the coupling of an extended mind, or are they merely constraints in the environment in which the agent acts? Presumably the relationship of use helps to distinguish between the agent and the environment. The agent comprises the coalition of users working together, together with the tools employed by those users in accomplishing the agent's task. Everything that is not

[41] Clark, A. (2002b) p 13-14

[42] See Barandiaran et al (2010), Bertschinger et al (2009) for more detailed definitions of agency and autonomy. This question is picked up again in Chapter 2.

immediately used in the process of completing the task is external not only to the user, but also *to the agent*, and can therefore be considered part of the environment in which the extended agent acts. If we accept the weaker view as I have been suggesting, the question of which systems are causally relevant to the behavior of the agent becomes an empirical (or phenomenological) issue of discovering what environmental resources are actually being used and which resources are not engaged in the moment of use.

One of the difficulties in drawing the distinction between agents and environments arises from the fact that (biologically) external resources are also common social resources. The signs and markings along a road are available to all the driving agents, and it is partly in virtue of this shared common resource that the agents can coordinate their activity. Manipulating common resources for coordinating the behavior among several agents is an old evolutionary trick called 'stigmergy', and is common throughout the animal kingdom[43]. This behavior is clearly observed in the foraging behavior of a colony of ants and architectural abilities of termites in constructing their nests.

In fact, Clark argues that natural language is exactly the kind of common external resource that human brains are designed to exploit in order to engage the world. Human brains have evolved the computational ability to tightly couple with linguistic artifacts it finds in its environment as it develops, and this tight coupling has (obviously) significant consequences for our ability to think about and engage the world. Language use is just a particularly special case of the more general ability to offload cognitive processes onto the local environment and recover the results—in other words, to use the world as a tool—and though human minds are not unique in this regard, our minds are especially suited to the task. Clark argues that language understood as an external resource is ultimately responsible for the uniquely human ability to "think about thinking", and all the cognitive and instrumental benefits that

[43] Clark, A. (1997) p 75-76. See also Gordon, (2010)

come with higher order thought. Thus, language is "the tool of tools":

> "[A]ssociating a perceptually simple, stable, external item (such as a word) with an idea, concept, or piece of knowledge effectively freezes the concept into a sort of cognitive building block—an item that can then be treated as a simple baseline feature for future episodes of thought, learning, and search… as soon as we formulate a thought in words (or on paper), it becomes an object for both ourselves and for others. As an object, it is the kind of thing we can have thoughts about. In creating the object, we need have no thoughts about thoughts—but once it is there, the opportunity immediately exists to attend to the object in its own right. The process of linguistic formulation thus creates the stable structure to which subsequent thinkings attach."[44]

The stability of linguistic items stored in the local environment is central to Clark's discussion of language. Graphemes stored with paper and ink, or phonemes stored in sound waves, are artifacts that are accessible by any system within the local environment geared to recognize and process those artifacts. The *perceptual* stability of these artifacts is understood relative to the systems poised to detect and manipulate language as an artifact. In other words, the power of language derives from its status as an external tool. Since linguistic artifacts are common environmental resources, any system that has the ability to detect and process these artifacts can work to coordinate their activity with other language users sharing this same environmental resource.

The case of language is instructive for thinking about the use relationship[45]. When two language users are engaged in conversation, these two users coordinate their activity via a shared common external resource of the language. Even if the agents are manipulating and processing these linguistic artifacts in distinct ways, they are both *using* the same language[46] as long as they are manipulating the

[44] Clark, A. (2001) Mindware. Oxford. p 144ff.

[45] There is an interesting contrast between Clark's discussion of words and the parallel discussion in Ebbs (1996; 2009). The contrast relates to the active/passive distinction Clark uses to distinguish his work from the Burge route of anti-individualist considerations of meaning and content; Ebb's suggestion that we take our ordinary use of words 'at face value' is analogous to Clark's phenomenological criteria for establishing parity among mind tools. This relationship may warrant further elaboration in future work.

[46] This is not to imply that the users mean the same things by the words they are using, or have the same level of understanding when they employ those words. If two systems are processing language in radically different ways, they may mean radically different things by their words. But as long as they are manipulating the same linguistic

same artifacts. Leaving the distinction between agents and environments messy allows Clark to describe this interaction in a variety of ways: there may be multiple agents engaged in the use of the same external artifacts, and therefore the two agents overlap in significant ways. Or there may be a single agent unified by these external objects, extended across multiple language users working together. Since agents are extended coalitions of users and the tools they employ, distinguishing between agents is partly an issue of how we want to divide up responsibility (credit and blame) for the completion of a task. I want to set this difficult issue aside, since it distracts from the core problem of the relationship between minds and machines.

We are now in a position to resolve the deep tension in Clark's treatment of the use relationship. A tool is an extension of the mind, and a machine becomes a tool by being used. However, because of the constraints Clark places on mind-extensions, the use of some machines cannot be accounted for by the user/tool divide. In other words, the use relationship is insufficient for describing the interactions between minds and certain kinds of machines. Even if the use of Google extends my agency in certain ways, my phenomenological and epistemological relationship to Google does not resemble my relationships with other tools. Since these interactions occur in the moment of coupling, the design relationship is also ruled out as a possible explanation[47]; and since systems such as Google aren't built to mirror any aspects of the human mind, treating Google as an artificially intelligent system in the sense of the classic AI debate is also ruled out as an option.

Reflections on the case of language use, however, reveal another possible interaction between minds and machines that is not available in the use, design, or mirroring relationships. When I engage another person in conversation, I am not 'using' that person in the sense in which I use a tool. Rather, I

---

artifacts, this process will have consequences any other system poised to detect and manipulate those artifacts.

[47] This step of the argument remains unsatisfactory to those who believe that the aspects of Google's behavior that cannot be explained by use can still be explained by design. Again, this step in the argument will be treated in more detail in the next chapter.

am cooperating with that other language user as a participant in a shared activity. In the remainder of this paper, I will argue that interacting with some machines can take the form of this cooperative, participatory relationship. However, in order to avoid the confusing relationship between language and intelligence, I'll turn to discuss the much simpler cases of interactions described by the perceptual crossing experiments.

2.5.1 Extended Agents and Participants in the Perceptual Crossing Experiments

Clark's extended mind hypothesis has recently found expression in an experimental paradigm within psychology known as "perceptual crossing", first proposed by Auvray et al (2006).[48] Auvray's 'interactionist' approach is inspired by Dennett, Clark, and others working within a framework of embodied, extended cognition, and will provide a useful framework for the discussion that follows. The experimental design provides a minimal framework for describing interactions between users and objects within a shared environment, will serve as grounding for the picture of machine participation defended in this thesis.

The perceptual crossing experiments are designed to examine the behavioral differences that arise from users engaging with agents and objects in their environment. To control the complexity of such interactions, the perceptual crossing experiments are designed to be as simple as possible. The entire experiment takes place within a one dimensional virtual space, essentially a continuous line which loops around on itself. Two users are each given control of virtual avatars that exists in the space; each can move their avatar left or right on the line with a track-ball they control with one hand. The two users exist on the same space, and therefore may cross each other's position in space; when this interaction happens, users are given haptic feedback in the form of a vibrating pad in their other hand. In this sense,

[48] See also Auvray et al. (2009). For discussion, see Rhode, M. (2008; 2010), Frosse et al (2014).

the users directly experience their interactions with each other in this virtual environment; hence, "perceptual crossing".

However, the virtual space contains more than the avatar of the experimental subjects. The space also contains a static object, fixed at one location on the virtual line, which results in haptic feedback when crossed. In this way, perception of the object is qualitatively identical to perception of the other user. Further complicating matters are two more objects, called 'lures', that trail the user's avatars by a fixed distance as they move around the one dimensional space, and that also provide haptic feedback when crossed. Since the distance is fixed, users can never encounter their own shadow. The result is that each user encounters a field inhabited by three different kinds of objects, one static and two moving. And one crucial difference distinguishes the two moving objects: when you cross the shadow, the other user feels nothing; but when you cross the other user's avatar, *they feel it too.* This distinction, of mutual recognition, makes all the difference.

Users are able to correctly identify each other with an incredibly high degree of accuracy, around 70%[49], a figure that may be surprising given the stark nature of the interaction and the feedback available. But more than the accuracy, results of the perceptual crossing experiments were interesting mostly due to how the identification challenge was solved. It is known from other experiments[50] that users interacting in minimal environments use a technique called *oscillation* to identify and describe the objects in their environment: basically, they wobble back and forth over the object, crossing it many times. These oscillations are indicative of agency: static objects don't oscillate at all; although lures move, they don't oscillate in response to being touched, because the user leading the lure doesn't experience the crossing. Only when the two user avatars come into contact do both users oscillate in

[49] See Auvray et al. (2009). Rhode (2008) and DiPaolo (2008) describe these experiments and their variations in more detail.
[50] See Sribunruangrit, N. (2004).

response to each other. It is this sustained interactive dance that distinguishes clearly the agent from the other objects (both moving and static) in the field[51].

Froese et al (2014) argue that these oscillating interactions themselves constitute the social cognition of agency, as users "negotiate the conventions of interaction", and suggest the experiments to be confirmation of the extended mind hypothesis and "participatory sense-making"[52]. Interestingly, their studies demonstrate that users spend far more time interacting with other users than they do with either the static objects or the lures (28996 avatar contacts compared to 5046 lure contacts and 5621 static object contacts)[53]. DiPaolo et al (2008) use agent-based modeling to confirm that the coordinated dynamical oscillation is an optimal strategy for discovering agency. Their models assume that an agent determines if it is worthwhile to oscillate back for further inspection, or to discard the experience as a static object, in terms of the total time spent interacting with the object. Since the static object and lure don't oscillate in return, interactions with them are discarded faster than interactions with other agents, resulting in a pattern of behavior that closely models the empirical results from perceptual crossing experiments.

The results are suggestive of a view of agency and participation as it relates to cognition and the extended mind, a view that will be elaborated in more detail in Chapter 3. Fundamentally, all machines are agents (even the static ones). Users are those 'special' agents whose functional task is to coordinate the activity of other agents in the environment. Some of the agents being coordinated are mere agents; these are the tool extensions of the coordinating user. But the participants are the agent in the environment that can sustain a participatory interaction with other agents in that environment. The sense of 'sustain' employed here is directly analogous to the sense in which a word can sustain the higher-order thoughts of a linguistic agent: it provides a stable, perceptible token upon which to build

[51] See Dennett (1989)
[52] Froese et al (2014). See also De Jaegher (2007)
[53] Froese et al (2014)

more abstract and elaborate conceptual resources. The perceptual crossing experimental paradigm provides a parallel analysis for participation: an agent is a participant if it provides the stable, perceptible experience through which other agents in the environment recognize it as an agent. In the minimal one dimensional environment, this amounts to displaying oscillating behavior in response to an encounter with another avatar. But we can imagine other, more elaborate environments, with more complex forms of interaction and participation that require sustaining. For instance, one might reasonably take the ability to sustain a conversation as a behavioral display sufficient for recognition as a participatory member of a linguistic community. In the next section, I'll argue that this is precisely the point of Turing's famous test.

2.6.1 Turing and participation

Approaching machines as participants in a shared, common activity is not unprecedented in the philosophy of mind. One of the first systematic philosophical approaches to the question of artificial intelligence discusses the relationship between humans and machines in exactly these terms. Turing's famous 1950 paper "Computing Machinery and Intelligence" finds the question "can machines think?" to be "too meaningless to deserve discussion."[54]However, we can come to discover that certain machines are intelligent, on Turing's view, upon engaging those machines in conversation and judging their behavior to be indistinguishable from the behavior of another human, what has come to be called the Turing Test. If the intelligence of the human is judged on the basis of their behavior, than an indistinguishable performance from a machine warrants the same judgment. Turing's 'test' for intelligence concedes a lot to those who doubt that machines can be genuinely intelligent. Appealing to conversational language use as a basis for judging intelligence presupposes that language use is a

[54] Turing, A.M. (1950).

paradigmatic example of intelligent human behavior, a suggestion first put forward by Descartes[55]. Perhaps this concession sets a poor standard for intelligence, since the ability to use language rests on more fundamental cognitive machinery[56] that can be mimicked with less sophisticated devices. Nevertheless, it sets the standard for intelligence incredibly high—so high, in fact, that machines still routinely fail to pass the test for linguistic competence nearly sixty years later[57].

Throughout his discussion of artificial intelligence Turing was extremely sensitive to address the bias he saw against machines. The ability to use language provided a means for offsetting this bias, since it drew a "fairly sharp line between the physical and intellectual capacities of man" and avoids being "weighted too heavily against the machine."[58] Turing's sensitivity to the bias against machines is surprising,[59] and is a sentiment not typically echoed in the classic debate described above. The classic debate provides what I'll call the "standard interpretation" of Turing's test[60], that it sets a criterion for machine intelligence: namely, behavioral indistinguishability from human conversational linguistic performances. The standard interpretation is that Turing believes if a machine meets that standard, then it should be treated as intelligent.

The standard interpretation of Turing basically stops there, although philosophical and technical questions still linger about the criteria Turing uses in the test. Most of the discussion about the test

[55] "... there are no men so dull and stupid, not even idiots, as to be incapable of joining together different words, and thereby constructing a declaration by which to make their thoughts understood; and that on the other hand, there is no other animal, however perfect or happily circumstanced, which can do the like… And this proves not only that the brutes have less reason than man, but that they have none at all." From Descartes' *Discourse*.

[56] Indeed, on Clark's view, our capacity for language rests on the more fundamental ability to exploit environmental resources as users.

[57] See Saygin (2003)

[58] Turing (1950) pg 343 emphasis added.

[59] … but not unprecedented. A defense of machines against human bias can be traced back to at least La Mettrie, who urges in his 1748 book *Man a Machine* to "break the chain of your prejudices, [and] render to nature the honor she deserves" by "conclud[ing] boldly that man is a machine."

[60] By "standard", I mean this aspect of the test is nearly universally agreed on and serves as the focus of discussion. The interpretation I'm presenting here doesn't disagree with the interpretation so much as argue that Turing's own emphasis was elsehwere. For an overview of the literature on the Turing test and the variety of interpretations it has generated, see: French, R. M. (2000).

involve various ways of extending the behavioral criteria into other domains of human competence (with the limit being the so-called Total Turing Test[61]), or more critically about whether behavioral tests are sufficient for identifying intelligence at all[62], but none of this has much to do with Turing's own views about machines. They are ways of using his criteria to address the question of thinking machines itself, and not attempts to address Turing's own complicated views of the machines themselves. In any case, we are left from the standard interpretation with the impression that Turing thinks intelligence is specific property (or set of properties), and that when machines come to have that property then the term should be applied to them.

By offering an interpretation of the test that emphasizes fair play, I'm not disagreeing with the standard interpretation of the test. Still, Turing's reasons for offering the test are obviously more complicated than the criteria itself suggests. For instance, the 1950 paper begins with a discussion of the traditional "imitation game", which was a parlor game where people try to guess the gender of the interlocutors. Turing explicitly uses this game as the inspiration for the Turing test; if we were to apply the standard interpretation to the original imitation game, it would imply Turing's views on gender were as follows: A man is anything that acts like a man, and a woman is anything that acts like a woman. I think it's reasonable to think that, given Turing's personal history, that his views on gender identity were a little more complicated than that. Rohde et al. (2010) explicitly interpret Turing as defending a purely ascriptional view of intelligence, and worry that it leaves room to wonder whether the ascriptions are accurate of their target. Rohde use the perceptual crossing experiments to suggest that social recognition requires not just the Turing-style ascription of intelligence, but also that the target itself behave in the appropriate way. They conclude that a Turing-style test for intelligence is inadequate to characterize its nature. As Dretske says, "…despite indistinguishability, all is dark… [the] Turing Test does

[61] See Harnad, S. (1992).
[62] See Searle (1980) and Dreyfus (1992), and the discussion from section 2 on mirroring.

not test for the right thing. It does not tell us what we want to know: whether the behavior… is for some purpose or… is explained by beliefs and desires."[63] On the assumption that "what we want to know" is some fact about the internal processing of the machine, the Turing Test fails to provide an adequate answer.

On my interpretation, Turing's views here are explicitly constructivist: there are no absolute criteria for gender identity (or intelligence), except what our collective social biases take them to be. However, this does not give Turing's view a completely ascriptional character. It matters deeply for Turing's test that the machine in question is capable of sustaining a convincing conversation. The whole point of the Turing test is to determine whether the computer "oscillates" as we would expect of a language user, such that other language users would recognize it as a participant in their conversations as they would any other user. Rohde's interpretation of Turing's test as in conflict with the results of the perceptual crossing experiments is fundamentally a failure to appreciate the ways in which our computer interlocutors (intelligent or not) participate in the imitation game.

The imitation game is designed to abstract away from a variety of uniquely human behaviors in order to evaluate the intelligence of the machine from as neutral and unbiased a perspective as possible, and many of the philosophers in the classic debate feel these abstractions don't do justice to the character of human thought. But Turing's concern throughout his discussion of thinking machines runs directly counter to these intuitions, and is instead motivated by a concern for "fair play for machines". Turing raises this concern in discussing the apparent infallibility of computing machinery:

> "… if a machine is made for this purpose it must in some cases fail to give an answer. On the other hand if a mathematician is confronted with such a problem he would search around and find new methods of proof, so that he ought to be able to reach a decision about any given formula. This would be the argument. **Against it I would say that fair play must be given to the machine.** Instead of it sometimes giving no answer we could

[63] Dretske (1994) p143, 159

> arrange that it gives occasional wrong answers. But the human mathematician would likewise make blunders when trying out new techniques. It is easy for us to regard these blunders as not counting and give him another chance, but the machine would probably be allowed no mercy. In other words, then, if a machine is expected to be infallible, it cannot also be intelligent."[64]

Turing's approach to the question of artificial intelligence is grounded in his concern for treating machines fairly in the context of a game, and this concern has obvious consequences for the nature and purpose of the imitation game. In this way, "fair play for machines" motivates Turing's entire discussion of machine intelligence; unfortunately, the consequences of adopting this principle are lost in the subsequent debate over whether machines can mirror the important structural features of the mind. Turing thinks the question "can machines think" can be answered in the affirmative without any deep understanding of how thinking or intelligence operate in brains; after all, I judge other humans to be thinking on the basis of behavior alone, and without knowing anything in particular about the way minds work. To evaluate a machine's intelligence by some other basis would violate the fairness principle. Fairness for the machine is clearly at the front of his thinking.

For more textual support for my interpretation, consider Turing's responses to objections in that 1950's paper, which repeatedly takes the form of "how might we apply the same standards of judgment to the machine?" Consider, for instance, the Theological objection Turing considers: god gave a soul to men and not to machines, therefore machines cannot think. His approach is not to argue that souls really are a part of the machine; instead, his response is to argue "why can't God give machines souls?" The strategy is as follows: If you think X is an important feature of human intelligence, then Turing will show that the very conditions under which you judge a human to have X can also be used to judge the machine to have X, and therefore we should conclude the machine to be intelligent. In the argument from consciousness, for instance, we consider humans intelligent because they are conscious, but

[64] Turing (1947)

insofar as we make that judgment on the basis of convincing linguistic performances then we might also make it of machines executing similarly convincing performances. If, on the other hand, recent science suggests some more precise performance is required for intelligence(for instance, the presence of certain specific patterns of neural activity), then again we'd want to judge the performance of the machine by fair standards. Again, Turing isn't arguing that consciousness or specific neural activity is essential to intelligence; he's arguing that *if we think it is*, then we should judge the machine by the same criteria we use to judge humans. This is fundamentally a plea for fairness for machines, not a defense of a strict ideological perspective on intelligence.

The upshot of all this is that Turing isn't arguing that "behavioral indistinguishability" is sufficient for intelligence, in the sense that it is an objective standard of judgment suitable for inquiry into the nature of the mind. Instead, Turing's test is meant to provide for the conditions under which a person might fairly judge the differences between a human and a machine without simply falling back on prejudice. Turing argues that on the basis of simple fairness, if this behavior is sufficient for judging the human intelligent, then it is also grounds for judging the machine to be the same. Turing isn't interested in defending (or refuting) the claim that humans are intelligent; he takes it for granted that we think they are, and his test is designed to allow us to apply the same judgment fairly to our machines.

Taking Turing's fairness principle seriously has radical consequences for the way we interpret the imitation game. For Turing's imitation game to get off the ground, it must be the case that I can play the game with any machine, whether or not that machine is a genuinely thinking mind. While Turing's plea for fairness is not an explicit endorsement of machine participants, it does suggest a kind of relationship we might have with machines that doesn't turn on our instrumental conception of artifacts, nor on the structures we happen to share with the machine. Moreover, it suggests that Turing's proposed test is concerned with our treatment of machines in the sense of their role in a social context

(like a game). In other words, while Turing's test is typically interpreted as addressing the issue of thinking machines, on my interpretation Turing is addressing an issue in the philosophy of technology. Specifically, he's proposing a way of thinking about machines not in terms of *kind* but in terms of *performance*, specifically in contexts where its performance is perceived as being an independent agent (as an interlocutor in a conversation). Insofar as this relationship doesn't fall under the instrumental conception of artifacts, Turing's proposal defies the dual natures view.

To be fair, this response only extends so far; what matters, how well the potential participant sustains the ascriptions I make. I might be able to pretend, for a moment, that a rock is playing chess, but I'm not going to get very good games out of a rock. The rock doesn't (typically) engage in behavior sufficient to sustain my pretension of its participation in the activity, at least for very long, if I'm serious about playing chess; eventually I'm bound to conclude that the pretension is futile. We might put this in terms of Ebb's technique[65] of face-value responses to the skeptic. Imagine I called you over to show you this amazing chess-playing rock. And say you're willing to suspend your immediate disbelief and evaluate the claim for yourself. But it's just a rock. It's not doing anything special; it just sits there across the board from me, quietly eroding. Even if you entered the situation with an open mind, after a few minutes watching the scene, any face-value judgment is bound to conclude that the rock simply isn't playing chess. By any ordinary judgment of typical chess-playing behavior, the rock is failing on all fronts.

Contrast the rock with Deep Blue, who not only issues chess moves in response to other chess moves, but whose moves are of genuine interest to other chess players as chess moves, and can even challenge the cognitive limits of the best experts in the field. Deep Blue sustains the perception of participation by the lights of ordinary judgment not just across one game but across tournaments; across history. The community of chess players recognizes Deep Blue as among the first great chess-playing computers; his status as a participant in the community has been sealed. Whereas the rock fails

[65] Ebbs (1996)

to sustain the attribution of agency by the lights of ordinary judgment, Deep Blue's behavior gives us plenty of reasons to maintain the attribution.

My interpretation of Turing is that this difference, of sustaining a perception of participation, is the underlying insight motivating his imitation game. A machine passes the test if it can encourage the sustained engagement from an interlocutor as a conversational participant over some duration of time, and fails if it can't. At the end of the conversation, the human judge is meant to issue what Ebb's would call a face-value judgment: were you talking with an intelligent creature? If so, then by the fairness principle we should consider the machine intelligent. Everything else about the test-- the linguistic framing, the blind judge, etc-- are all designed to help the interlocutor make a face-value judgment that is not biased by prejudices against the machine that are irrelevant to the evaluation being considered. In other words, they are motivated by fairness.

This is not the standard interpretation of the test. The standard, cartoon interpretation is that Turing thinks "acts like a human" is synonymous with "is an intelligent creature", and the Turing test is a way of determining which machines act like humans. This is the purely ascriptional view attributed to Turing by Rohde and Dretske. On this view Turing is committed to a substantive theory of mind and draws a line in the sand to demarcate the machines that qualify as intelligent and those that don't, and the worry is that Turing has drawn the line in the wrong place. It's not particularly novel for me to argue that Turing doesn't really care about the mind in the way the cartoon interpretation suggests; he says it himself explicitly plenty of times. But my interpretation is that Turing's view turns on a general attitude towards machines that he suggests we take *regardless* of their intelligence. I'm not aware of any interpretation of Turing that views the test as an explicit philosophy of technology.

Deep Blue doesn't pass the conversational language test, so doesn't pass Turing's test for intelligence. But that doesn't prevent Deep Blue from being included in the participatory community of chess players. Its chess-playing behavior is sufficient for any other player to be considered among the

participants in the community. To deny Deep Blue its status as a participant because of its lack of intelligence would be a violation of fairness. This is fairness owed to Deep Blue not as a result of its intelligence, but its mere membership as a community participant.

The point for the purposes of this chapter is to establish that Turing's imitation game rests on the assumption that humans can engage in participatory interactions with machines. Participating with a machine does not require that the machine has a mind, and it doesn't rest on any similarities between the internal operations of humans and computing machines. Granting the machine participant status simply rests on a basic plea for fairness. When I am engaged in certain kinds of tasks with a computer, I do not need to treat the machine as a mere tool to be used. Depending on the context, I can take the machine to be a genuine opponent, contributor, or participant that I work with in order to accomplish the task. This is most clearly seen in the context of a game, where I treat the machine as a competitor and opponent, but is also required for treating the machine as an interlocutor in a simple conversation. The importance of introducing machines as participants in a shared, social activity is central to Turing's conception of thinking machines, and follows straightforwardly from the fairness principle.

Furthermore, treating machines as *participants* in a shared activity provides one possible solution to the *engineering* problem of actually constructing machines that can meet the criteria of indistinguishability. Turing's suggested solution is to build machines that can learn and adapt to its environment. Insofar as the appropriate use of linguistic conventions indicate intelligent behavior, it is necessary to introduce learning machines into a rich social environment where they can interact with other users of the same conventional language.

> "The machine must be allowed to have contact with human beings in order that it may adapt itself to their standards. The game of chess may perhaps be rather suitable for this purpose, as the moves of the machine's opponent will

automatically provide this contact."[66]

Formal games like chess allow machines to participate in social activities with humans in a straightforward and well-defined manner, and Turing believes that even this sort of context-independent interaction is sufficient for building machines that can adapt to human standards. Building a machine that is a competent conversationalist in robust and context-dependent settings adds the rhetorical punch necessary dissolve our bias against the possibility of thinking machines, but is nothing more than a compelling intuition pump; Turing readily admits that he has no convincing positive argument in favor of the fairness principle, yet it is a principle we must adopt if we ever hope to have genuinely intelligent machines.

2.6.2 Google is a language user

Clark requires a way of describing our relationship to machines that does not treat them simply as tools, since certain systems like Google fail to meet the phenomenological and epistemological constraints on mind extensions. Since my interaction with the search engine fails the criteria of immediate endorsement, the relationship between Google and my mind is more like my relationship with *other minds* than Otto's relationship with his notebook. Consequently, Clark's narrow phenomenological description of tool use would be rendered deeply inadequate for understanding Google as a system. The difficulty of the phenomenological approach in describing technologies like Google is worse still—not only is Google not an extension of my mind, Google isn't an extension of anyone's mind! The search results offered by Google do not reflect the thoughts and mental attributes of Google's designers any more than they reflect my own mind. One of Google's designers might be just as surprised as I am by the search results Google offers. Although the designers might be in a unique place to illuminate the internal computational and algorithmic operations of Google as a system, how

[66] *Ibid*

Google is used and how that use transforms the linguistic environment is anyone's guess.

If Clark is right to suggest that language is best understood as a shared external artifact poised for easy manipulation by users, then it is important to keep track of which users are actively engaged in manipulating those artifacts to determine who is within the community of language users. In just this sense, Google (the search engine) is itself a language user, manipulating and coordinating the deployment of linguistic artifacts independent of my purposes as its user. Google is poised to recognize the same linguistic artifacts that other members of its language community use, and the effects of its manipulation of those artifacts are made public for other language users in ways that have consequences both for their behavior and for the development of the language itself. Put simply, Google uses language in a way that matters for the way we all use language. Google's sensitivity to subtle changes in the language, and the resulting transformation of the linguistic environment we all share, normally occurs without the foreknowledge or consent of any other agent, extended or otherwise. This independent responsiveness to its (linguistic) environment may not make the system a mind, but it certainly makes it a relevant participant in the activity.

The point is that a search engine like Google cannot simply be classified as a tool. For instance, compare a search engine to a telescope. A telescope is a paradigm tool used for looking at the stars. The telescope receives electromagnetic waves as input, does some optical processing of the light, and feeds the resulting information to its user. The user is responsible for pointing the telescope in the right direction and focusing its lenses to satisfy the user's interests. There are similarities between Google and a telescope. You might think of Google as a way of 'looking' at the internet. If you want to find something in internet space, you point your search engine in the direction of the thing you are looking for. "Pointing" the search engine is accomplished by feeding the system a string of words, which can be refined and focused to narrow your search. Certainly we use Google to look at the internet, in the same

basic way that we use a telescope to look at the stars.

But Google is not just a tool, in the sense that it performs some set of operations that extend the users capacity. In order to generate a response to a search query, Google must spend time crawling the internet, looking for new links and correlating those with the other links it monitors in its database. It must also be able to evaluate a query in order to determine which responses are appropriate, and to adjust its standards for appropriate on the basis of the query and response behavior it observes in its users. The characteristic relationships between minds and machines discussed above are entirely inadequate for discussing our relationship with these systems. My ability to use Google as a tool for looking at the internet rests on Google's own ability to use language, and Google's status as an independent user cannot be explained in terms of Google's relationship to my mind. Clark himself argues that there is an epistemological and phenomenological gap between the mental operations of the user and the operations of Google, and for Google to count as a genuine tool this gap would need to be closed. However, rendering Google more transparent to the operations of my own mind would hamper its contributions to the task which I employ it to solve. Appealing to the original design of the system is unhelpful, since as Turing suggested it is necessary to introduce Google to the rich linguistic environment of the internet that allows the system to conform to the standards of its linguistic community. The applications to which Google can be put are partly the result of what it has learned over the history of this exposure, a history that was not anticipated in the development of Google's original design. Finally, no one thinks that Google's manipulation of linguistic artifacts mirrors the processing of language accomplished by the brain. Google is not a *natural language* user. It simply operates on the linguistic artifacts we make available. It is cooperating with us in a very literal sense, and by Turing's fairness principle can be taken as a participant in the linguistic community whether or not its operations mirror the linguistic operations in our own minds. Google is a machine participant that sustains our practice of attributing participatory agency to it, and in doing so makes a difference to the way our

sociotechnical community develops. I'm not arguing that this is sufficient for attributing intentional mental states to Google; on the contrary, I'm arguing that intentional mental states don't really matter to the question of participatory agency, social organization, and technological development. Hence, "artificial intelligence beyond the philosophy of mind". I take this to be Clark's explicit view, and Turing's view as well, although the latter has been underappreciated. I also take it to be the correct view of sociotechnical development, and one worth defending.

While Clark's discussion of the extended mind is not designed to address the existence of these quasi-independent systems, and I have tried to argue that his discussion is fundamentally under-equipped for even describing such systems, it does provide the basic conceptual resources for offering a new and relevant definition of thinking machines relevant to the discussion of artifacts that began our inquiry. Thinking machines are those technological systems that coordinate their processing with other active agents in participatory relationships, and do not function as an extension of any other agent as a tool. Artificially intelligent systems are agents and users in their own right, employing tools to conduct meaningful, intelligent work on the environment they happen to share with extended human agents. This definition retains the justification for philosophy of mind's interest in artificial intelligence, but shows that a full discussion of AI cannot be exhausted in the philosophy of mind alone. An artificially intelligent system might have a mind like ours, or it might have certain structural features in common with our minds, but these requirements are by no means necessary. Labeling participatory machines "artificial intelligence" seems appropriate on the definitions I've developed here, and although it is not the sense of the term typically given in philosophy I've argued that it is relevant for reading Turing and, I think, important for understanding the actual systems generated by artificial intelligence research.

It is reasonable to still worry that the issue of design hasn't been adequately treated. Maybe Google can't be understood simply as a tool, but perhaps its other aspects are accounted for by their

design (a distinct kind of mind-relation), and that the combination of both design and use can account for the behavior of the machine. Turing addresses this worry explicitly in his essay, in his discussion of Lady Lovelace's objection. This will be the focus of the next chapter.

Chapter 3

Lady Lovelace and the Autonomy of Machines

3.1 Abstract

Ada Lovelace claims that machines can only do what we know how to order it to perform. I treat the Lovelace's objection as a clear and strong statement of skepticism about machine autonomy. I begin by distinguishing the Lovelace objection from the general skepticism of thinking machines as it arose in the classic debates over artificial intelligence. I argue that recent attempts to discuss the problem of autonomy prove inadequate for responding to the strong skepticism of the Lovelace objection. I conclude by revisiting Turing's original response to the objection, which under my interpretation provides an adequate response to Lovelace that has been underappreciated in the classic debate. A proper understanding of Turing's positive account will, I claim, provide a general framework for thinking about autonomy that is appropriate for productively discussing the social situated, participatory machines that inhabit our contemporary technological world.

3.2 Taking autonomous machines seriously

According to the US Department of Defense, as of October 2008 unmanned aircraft have flown over 500,000 hours and unmanned ground vehicles have conducted over 30,000 missions in support of troops in Iraq and Afghanistan[67]. Over the past few years a number of government and military agencies, professional societies, and ethics boards have released reports suggesting policies and ethical guidelines for designing and employing autonomous war machines[68]. In these reports, the word

[67] USDOD (2009) Unmanned Systems Integrated Roadmap

[68] In addition to the USDOD roadmap above, the US National Research Council (2005) has advised the US Navy to aggressively pursue the use of autonomous vehicles. The European Robotics Research Network compiled a Roboethics Roadmap, see Veruggio (2006). The South Korean Ministry of Commerce, Industry and Energy produced a Robot Ethics Charter in 2007; see Shim (2007). The Japanese government assembled a Robot Policy Committee in 2008, see Salvini et al. (2010). Noel Sharkey has been particularly vocal about the importance and increasing urgency of robot ethics. See Sharkey (2008).

'autonomous' is used more or less uncritically to refer to a variety of technologies, including automated control systems, unmanned teleoperated vehicles, and fully autonomous robots. Describing such artifacts as 'autonomous' is meant to highlight a measure of independence from their human designers and operators. However, the very idea of autonomous artifacts is suspiciously paradoxical, and little philosophical work has been done to provide a general account of machine autonomy that is sensitive to both philosophical concerns and the current state of technological development.[69] Without a framework for understanding the role human designers and operators play in the behavior of autonomous machines, the legal, ethical, and metaphysical questions that arise from their use will remain murky.

My project in this chapter is to build a framework for thinking about autonomous machines that can systematically account for the range of behavior and relative dependence on humanity of our best machines... In chapter 1 I looked at the relation between a user and a tool in the moment of use; in this chapter I focus instead on the relation between the designer and the tool, and its relevance for thinking about autonomous machines. Pursuing this project requires that we take autonomous machines seriously and not treat them as wide-eyed speculative fictions. My concern is not with distant future science fiction possibilities, but with the live challenges concerning machines that are available and in use today. As a philosophical project, taking autonomous machines seriously requires addressing the skeptic, who unfortunately occupies a majority position with respect to technological artifacts. The skeptic of machine autonomy holds that any technological machine designed, built, and operated by human beings is *dependent* on its human counterparts in a way that fundamentally constrains its possibilities for freedom and autonomy in all but the most trivial senses.

In this chapter I respond to the skeptic in order to clear the ground for an account of machine

[69] Notable attempts include Haselager (2005) "Robotics, Philosophy, and the Problems of Autonomy", and Haselager and Gonzalez eds. (2007). The issue of autonomy is indirectly addressed in the literature on machine *agency*, and in particular the idea of a 'software agent'. See Stuart (2002). For a discussion of autonomy in information systems generally, see Bertschinger et al (2008), Barandiaran (2009), and Tani (2009).

participation. I will treat the Lovelace objection, cited in Turing's famous 1950 discussion of thinking machines, as the clearest and strongest statement of machine autonomy skepticism (MAS). I argue that the Lovelace objection is best understood as a version of the dual natures theory of artifacts (DNTA). Thus, a rejection of DNTA entails a rejection of MAS. In section 3 I survey some recent attempts to discuss the problem of autonomy as it arises in the robotics literature, and I argue that these treatments fail to adequately address the theory of artifacts that grounds Lovelace's objection. In section 4 I argue that the received interpretation of the Lovelace objection, which treats machine autonomy as an epistemological issue, likewise misses the force of her radical skeptical argument. I then argue that Turing's original response to the Lovelace objection is best understood as an argument against DNAT, and provides a uniquely adequate answer to the skeptic. On my interpretation, Turing's positive account of "thinking machines" provides a framework for thinking about autonomous machines that generates practical, empirical methods for discussing the machines that inhabit our technological world. I conclude by arguing that developments within computer science focused on "machine learning" vindicate Turing's account of the social integration of humans and machines.

3.3 Machine Autonomy Skepticism (MAS)

According to the dual natures theory of artifacts, artifacts depend for their natures on human mental activities. While explicit statements of this view are rare in the artificial intelligence literature, they pervade recent metaphysics, philosophy of mind, and philosophy of technology. Consider, for instance, the following formulations of what appear to be the consensus view: Baker (2006): "Unlike natural objects, artefacts have natures, essences, that depend on mental activity"; Kroes and Meijers (2006): "Physical objects, with the exclusion of biological entities, have, as physical objects, no function and exhibit no 'for-ness': they acquire a teleological element and become technical artefacts only in relation to human intentionality"; McLaughlin (2001): "The function of an artefact is derivative from the purpose of some agent in making or appropriating the object; it is conferred on the object by the desires

and beliefs of an agent. No agent, no purpose, no function"; Emmeche (2007): "Machines are extensions of human capacities and intrinsic parts of human sociocultural systems; they do not originate ex nihilo or fall readymade from heaven"; see also Vaccari (2013).

Machine Autonomy Skepticism finds a clear home in DNTA. MAS can appeal to DNTA to argue that even the most sophisticated performances of machines ultimately depend on human intentions and control. This is true even in cases where the machine appears to be engaged in independent performances, such as a self-driving vehicle, since these performances betray the careful human design that makes the performances possible. Since all so-called 'autonomous machines' are artifacts, machine autonomy skepticism concludes that there are no genuinely autonomous machines.

DNTA is invariant with respect to technological advance, since any potential technological innovation would presumably continue to involve the production of artifacts of this dependent sort. MAS inherits this invariance as well. Even in distant future cases where our machines are entirely designed and produced by other machines, the automated manufacturing process presumably still began as a designed, engineered human artifact at some nascent stage. Thus, the results of these processes necessarily bear the mark, however distant, of human mental activity, making them necessarily artifacts[70]. To suggest otherwise is to suggest that artifacts fall ready-made from the sky--precisely the claim rejected by the DNTA.

Besides its intuitive appeal, MAS has a strong rhetorical advantage in the debate over autonomous machines. Since MAS asserts that there are no autonomous machines, the use of unmanned drones and other advanced technologies poses no special ethical, legal, or metaphysical problem deserving of unique consideration. Instead, such devices ought to be treated like any other

[70] In this way, human design functions conceptually like the "original sin" for artifacts. This argument for DNTA is structurally analogous to the cosmological argument, in that it proceeds by tracing an effect to some original cause of a categorically different sort. This suggests to me that the formal structure of these causal arguments are deeply embedded in our intuitive judgments of the world. Therefore, one open line of attack on MAS would be to question the reliability of these intuitive judgments concerning causal source. In this essay, I've decided not to attack the psychological plausibility of the thesis, but to go directly after the metaphysical thesis itself.

tool, where the operator or designer is attributed with, and ultimately held responsible for, the behavior of the machine. With certain technologies there may be a question of how to divide the responsibility among a collection of designers and operators, but worries about collective responsibility are again not unique to the domain of the technological, thus so-called "autonomous machines" still pose no special problem. Maintaining a categorical distinction between artifacts and minds has the rhetorical advantage of clarity and metaphysical simplicity that apparently projects into the distant future and across radical technological change. This gives MAS the appearance of epistemological stability and parsimony, both of which can reasonably be considered methodological virtues of the position.

These methodological virtues are not unique to the skeptic of machine autonomy. One alternative position holds that some future technological advance may make autonomous machines possible, and in the event they are created ought to be treated similarly to any other autonomous agent, i.e., as a person, with all the traditional legal, ethical, and metaphysical privileges this status entails[71]. On this view, autonomous machines pose no special philosophical problem, because they simply raise the familiar issues of autonomy as it pertains to all other persons, human or otherwise. This view implicitly rejects DNTA, since it explicitly accepts that autonomous machines may be possible. However, it would also reject the proposal that unmanned self-guiding drones are autonomous, since they are clearly *not* persons in the relevant sense comparable to human agency. I feel this response fails to take machine autonomy seriously as a live challenge, entertaining it merely as a logical possibility, and is therefore not the target of my argument in this chapter. The skeptic I'm interested doesn't merely argue that no machine available today is autonomous; MAS is the view the no artifact can in principle be consider autonomous. Nevertheless, a rejection of MAS has implications for views that identify autonomy and personhood, and I'll return to consider those implications at the end of the chapter.

Machine autonomy skepticism is most forcefully stated by Ada Lovelace, as quoted in Turing

[71] Thanks to Helga Varden for pointing this out.

(1950): "'The [machine] has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform' (her italics)." If the machine's performance is not original but derives its orders from humans, then the machine does not enjoy the requisite independence to be considered autonomous. Lovelace's objection might be made explicit as follows:

> **L1:** Machines can only do whatever we order them to perform.
> **L2:** Therefore, machine performances are not original, but derive from human orders.
> **L3:** Therefore, machines are not autonomous.

This argument appears compelling across a variety of positive accounts of autonomy, which will be treated more carefully in the next section. Some particular account of autonomy is necessary to treat the move from **L2** to **L3**. My goal in addressing the skeptic is not to defend (or attack) any particular account of autonomy, but rather to undermine the categorical restriction on machines that motivates **L1**, and the inference from **L1** to **L2** above, which will be common to any MAS account regardless of their particular views on autonomy.

Distinguishing the skeptic of machine autonomy from the traditional skepticism of artificial intelligence requires situating the latter form of skepticism within what I call the classic philosophical discussion of artificial intelligence, which occupied philosophers from Turing's paper to the end of the twentieth century[72]. Central to this debate was an incredibly compelling analogy between the paradigms of human intelligence and the formal operations of computers. As Dennett says,

> "Computers are mindlike in ways that no earlier artifacts were: they can control processes that perform tasks that call for discrimination, inference, memory, judgment, anticipation; they are generators of new knowledge, finders of patterns–in poetry, astronomy, and mathematics, for instance–that heretofore only human beings could even hope to find."[73]

Beyond the analogy, computers appeared to solve a problem that had plagued philosophy since we began working out modern materialist science four hundred years ago: the relationship between mind and matter. Computers demonstrated that it was possible to have complex, intelligent, meaningful

[72] This debate is covered more thoroughly in Chapter 1.
[73] Dennett, (1999)

activity carried out by a purely mechanical system. Dennett continues:

> "The sheer existence of computers has provided an existence proof of undeniable influence: there are mechanisms–brute, unmysterious mechanisms operating according to routinely well-understood physical principles–that have many of the competences heretofore assigned only to minds."[74]

This existence proof not only provided philosophers with a satisfying demonstration of how mental processes could be realized in physical matter, but has also generated a practical, empirical method for studying the mind: to understand how the mind performs some task, build an artificial system that performs the task and compare its performance with our own. Thus, we were able to turn an abstract conceptual problem into a positive scientific research agenda, and the force of the old philosophical puzzles of the mind began to dissolve as scientists took the reins.

The classical skeptics of artificial intelligence, like Dreyfus (1972, 1992) and Searle (1980) remained unimpressed by this analogy, and proposed a variety of activities that computing machines simply could not do: computers did not understand the meaning of sentences or significance of the tasks they performed, they had no conscious access to their environment or their internal states, they did not approach the world with the emotion or affect that is characteristic of our merely human natures[75]. Central to their attacks was an emphasis on the relevant disanalogies between the operations of formal computing machines and the complex behaviors of dynamic, embodied biological organisms. Partially due to the incredibly difficult task of clarifying exactly what was missing from our best machines, and partially due to the plodding progress in computing machinery and cognitive science, the artificial intelligence debate stalemated in the early 90s with both camps claiming a de facto victory and little consensus achieved. There has since been relatively little movement in the philosophical debate despite the terrific advances within cognitive science and other AI-related fields[76]. Today, the issue of

[74] *Ibid.*
[75] See Haugeland (1981).
[76] These advances include, but are certainly not limited to, machine learning, neural networks and computational brain theory, social and evolutionary robotics, ALife, etc. See Mira (2008).

artificial intelligence is a mature research program in the cognitive and computer sciences, and is more widely recognized in philosophy as a pedagogically useful thought experiment raised briefly in introductory classes than as a live philosophical challenge.

Turing's presentation of the issue in 1950 includes a number of objections to the possibility of artificial intelligence, much of which anticipates the moves that were more thoroughly explored in that classic debate. But Lovelace's objection is unique among the objections Turing considers, and is explicitly singled out for a second treatment and response in the concluding remarks of that famous paper. Implicit in the Lovelace's objection is a fundamental concession to the proponents of artificial intelligence that restructures the nature of the debate and motivates Turing's positive suggestions in that final section. Specifically, Lovelace's objection does not entail that any particular feature of the mind is in principle off limits to a machine qua machine; c*ontra* the classic AI skeptics, Lovelace argues that the performances of the machine are limited only by the creativity and ingenuity of its human designers and operators. For this reason, Lovelace concedes a broad range of possible computer performances, including those that may appear to be uniquely human. For instance, even if "understanding" is properly analyzed at the level of the causal interactions between neurons, Lovelace can fully admit that this same structure can be realized by a computer simulation or some other artifact designed by appropriately clever scientists and engineers, such that it would count as a mechanical performance of that very same mental process. Nevertheless, Lovelace contends that the machine is only capable of such performances because of the "ordering" performed by those designers, and its performance is therefore fundamentally dependent on human input and control.

Put simply, the Lovelace objection is not an argument that machines cannot have minds. In contrast to every other objection Turing considers, and in contrast to the most prominent AI skeptics from the classic debate, Lovelace's objection does not turn on denying any particular performative activity of the machine, or on identifying some dissimilarity between that performance and its human

counterpart. This presents a unique challenge to Turing's imitation game and its behaviorist motivations. Compare, for instance, Turing's objection to the argument from consciousness. Professor Jefferson argues that machine performances don't issue from "thoughts and emotions felt", and therefore lack the consciousness characteristic of human thinking. Turing replies with the challenge to distinguish between those performances that issue from consciousness and those that don't in the context of an instance of the imitation game. Insofar as Jefferson maintains that the performances themselves are distinct, then examples of indistinguishable performances would yield a satisfying answer to Jefferson's objection. Similarly, Searle's Chinese room example argues that a behaviorally indistinguishable linguistic performance by a digital computer nevertheless fails to reproduce an important feature of human thought, namely its capacity for understanding, thus rendering the computer's performance fundamentally distinct from its human counterpart. Lovelace's objection, in contrast, maintains that the distinction lies not in the quality of the performance, but rather from the nature of its prior design. Even if the machine can engage in an indistinguishable performance, it would still nevertheless depend on human activity in virtue of its status as a designed artifact. The presence of human design provides a clear basis for distinction that Lovelace argues isn't muted by the machine's eventual performances. Thus, Lovelace argues that technical artifacts cannot perform independently of our orders; her objection presents a skeptical argument against the very possibility of machine autonomy due to the influence and orders of our design.

This focus on the dependence of machines on human activity reorients the debate over artificial intelligence in two important ways. First, Lovelace's objection does not necessarily endorse any particular view of the operations of the mind, and therefore her objection is out of place in the debate over the foundations of cognitive science, discussed in Chapter 1 as the "mirroring" relationship, where much of the classic artificial intelligence debate took place[77]. Like Turing's own position, Lovelace's

[77] See Haugeland (1981), and the discussion in Chapter 1.

objection obviates the need to provide any criteria or working definition of thinking (or any other cognitive or mental process) in order to evaluate the performances of the machine. Lovelace is not arguing about the limitations of the formal processes of computation to render mental processes, or the ineffable nature of mental activity itself, but instead about the unique dependence of machines on humanity, and its implications for any potential machine performance. While the classic skeptic is interested in the nature of the mind, and therefore considers the technology as such irrelevant[78] to this investigation, Lovelace appeals to a more general theory about the nature of technology[79] that outstrips the local concerns about the nature of minds. Placing the debate outside the framework of a theory of mind sets the Lovelace objection apart from all forms of classic skepticism.

This change of focus implies a second, and perhaps more important, difference between Lovelace and traditional AI skeptics. Lovelace's emphasis on machine autonomy undermines the importance of the analogy between humans and computers central to the classic debate. Since Lovelace appeals directly to the relationship between intelligent designers and the artifacts they design, her objection targets a broader class of technical artifacts than just those claimed to be mind-like due to some structural or behavioral similarity between their performances and those of a human. No one mistakes unmanned military drones or the Mars rovers for human analogs, or think they have 'minds' in anything but the most philosophically controversial senses. Their purported autonomy does not rest on an analogy with human minds, but rather on the relationship between their behavior and the

[78] *ibid,* especially the technological irrelevancy thesis: "The crucial issue is not protoplasm versus semiconductor ("wetware" versus "hardware"), but rather whether the product is designed and specified in terms of a computational structure. If it is, then a working model could probably be manufactured more easily my means of electronics and programming; and that's the *only* relevance of the technology." (Haugeland, 1980, italics original). Again, see Chapter 1.

[79] This more general theory of technology does find some purchase in the classic debate, for instance in Searle's (1980) infamous distinction between 'strong' and 'weak' AI, which revolves precisely around treating a machine as a tool and treating it as a mind.. Searle's distinction is grounded in precisely the same conception of technology at work in Lovelace's objection. On this view, tools, unlike minds, cannot be understood independent of their purposive design and operation. However, Searle's interests in that paper are not in the nature of tools and the technological; as a classical AI skeptic, he is interested in the nature of thinking, and so he leaves this technological distinction unanalyzed. See the discussion of technology in Chapter 1.

intervention (or lack thereof) of their designers and operators. These machines invite the challenge of Lovelace's objection quite independent of the issues central to the classic debate.

Together, these differences make clear that Lovelace's argument for machine autonomy skepticism is best understood from the perspective of a "dual natures" view of artifacts: no possible performance of an artifact is sufficient for undermining its nature as an artifact. Artifacts by their very nature depend on the mental activity of their designers, and are therefore are not candidates for autonomy even in principle. While the DNTA was often left implicit in the classic debate, I suspect that it is the source of much of the hard-nosed skepticism and chauvinism[80] that, in part, gave rise to the eventual stalemate in the classic debate. Making this assumption explicit suggests that a satisfying answer to the skeptic cannot depend on an analogy between the performances of the machine and their human counterparts, and therefore will not mirror the familiar arguments from the classic debate. Instead, developing a satisfying account of machine autonomy requires an explicit analysis of the relationships between artifacts and their designers and operators. I turn in the next section to analyze this relationship.

3.4 Varieties of Autonomy

The very idea of machine autonomy strikes many philosophers as a nonstarter. Muntean and Wright (2007) identify an intuitive fundamental incompatibility between the concepts of 'mechanism' and 'autonomy'. Bickhard (2007) argues that the very idea of 'mechanism' is insufficient for capturing the normativity of representation and cognition required for autonomously minded systems. Haselager (2005) identifies three different conceptions of 'autonomy' that often run together in the discussion of

[80] Dennett (1998): "Let us dub *origin chauvinism* the category of view that holds out for some mystical difference (a difference of value, typically) due *simply* to a fact about origin." Dennett characterizes this view under the heading "Consciousness abhors an artifact", and claims it is "deserving of our ridicule because if taken seriously it might seem to lend credibility to the racist drivel with which is shares a bogus intuition." (p 156ff)

autonomous machines. In this section I'll review Haselager's discussion in order to further elucidate Lovelace's objection.

The first of Haselager's conceptions of autonomy is active within the robotics literature and deals primarily with mechanical *automation*, while the second is part of a much older and staid philosophical tradition that might be dubbed *genuine autonomy*. Haselager argues that these two conceptions are fundamentally irreconcilable; however, he goes on to advocate a third, holistic conception of autonomy that might bring the two opposing discussions into dialogue. I will treat each of these conceptions in turn. I will argue that, although superficially similar, Lovelace's strong skepticism is not tied to the traditional conception of autonomy, and is best understood as rejecting a key assumption within the robotics literature. This will show Lovelace's objection to be a particularly radical form of skepticism that threatens the possibility of machine autonomy in all three of Haselager's proposed senses.

Haselager introduces the first conception of autonomy by noting that robotic agents have attracted significant interest in recent work on artificial intelligence and seem likely candidates for autonomy simply in virtue of their "capacity to 'do something'"[81] without immediate human control. Haselager traces the development of the concept of autonomy in contemporary robotics to industrial uses of remotely controlled or 'teleoperated' systems in the hazardous environments associated with nuclear research in the late 1940s. Like the unmanned military vehicles in use today, early teleoperated systems automated certain basic mechanical functions, thus relegating its human operator to the merely 'supervisory' role of remotely monitoring and coordinating the performance of those functions. Describing such systems as 'autonomous' naturally grew out of this increased automation. "Increasing the autonomy of the remote simply means reducing the need for human supervision and intervention. Autonomy is interpreted relative to the amount of on-line (while the robot is operating) involvement of

[81] Haselager (2005) p 516. Cf my definition of machines in chapter 1 as "artifacts that *do something"*.

human operators."[82] Drawing from Franklin and Graesser's (1997) survey of the artificial intelligence literature, Haselager formulates a definition of autonomy designed to capture this use of the term. For clarity's sake, I will refer to this definition as 'automaticity'[83] to respect the clear emphasis on *automation* (in the literal sense of 'self-moved'[84]) in contrast to the more familiar notion of *autonomy* (in the sense of 'self-governed') discussed below.

> **Automaticity**: **"**An autonomous system is thought to be able to operate under all reasonable conditions without recourse to an outside designer, operator or controller, while handling unpredictable events in an environment or niche."[85]

A number of features of Haselager's definition of autonomy as automaticity are worth emphasizing. First, automaticity is relative to the intervention of human operators in the moment of the machine's performance, and thus best understood in terms of the *use* relation as discussed in chapter 1. Haselager explains: "The 'without recourse to an outside designer' refers to recourse during the act (i.e., on-line), but not to recourse to a designer preceding the behavior (i.e., programming)."[86] Distinguishing between online and offline behavior likewise demarcates when the interference of designers is relevant or irrelevant to its operation as an autonomous system. Slamming on the brakes while a self-driving car is operating interferes with its autonomy, but coding a breaking routine into its software before turning the car's autopilot on does not. Second, automaticity is a matter of degree. A system can be more or less automated depending on the amount of influence its human operators have in the moment of performance. Finally, Haselager explicitly concedes that this definition of autonomy "lacks philosophical

---

[82] Haselager (2005) p 518

[83] Thanks to Ty Fagan for this suggestion.

[84] It is worth noting that the history of philosophy has long recognized the automation of all living things, in the sense of Aristotle's discussion of 'animate' creatures in *De Anima*, and thus the focus on automaticity threatens to reintroduce the seductive analogy dismissed in the previous section. For instance, Gehlen (1957): "And in fact, in a number of quite central aspects of his own nature man himself is an automatism: he is heartbeat and breath, he lives in and by a number of meaningful, functioning, rhythmical automatisms-- think of the motions of walking, think above all the ways in which the hand operates. Think of the "circle of action" which goes through object, eye, and hand, and which in returning to the object concludes itself and begins anew."

[85] *Ibid.*

[86] Haselager (2005) p 518

import."[87] Specifically, automaticity refers to the relative amount of independence a system has in pursuit of its goals, but genuine autonomy in the strong philosophical sense also requires the agent to be actively involved in setting the goals it pursues, and it is only in this latter sense that the dual natures account will begin to object. The offline design prior to the machine's performance often includes the explicit specification of not only a machine's goals, but also the methods for identifying and accomplishing those goals; this activity is not typically attributed to the machine itself, regardless of its eventual online activity. In other words, this offline activity (what is typically referred to as "programming", and what Lovelace called "orders", whether or not it involves coding in the technical sense) is precisely what renders the object an artifact of its designers, in the sense that motivates the DNTA discussed in the previous section. For this reason, Haselager acknowledges that the roboticist's definition of automaticity will strike philosophers as unsatisfactory for capturing true autonomy.

Haselager's second conception, which I call 'genuine autonomy', aims to capture this stronger philosophical conception:

> **Genuine Autonomy: "**Autonomy is deeply connected to the capacity to act on one's own behalf and make one's own choices, instead of following goals set by other agents. The importance of being able to select one's own goals is also part and parcel of the common sense interpretation of autonomy."[88]

This conception of autonomy is deeply tied to issues of the freedom of the will and other well-worn philosophical conundrums, and Haselager notes that ""it would not be unreasonable for roboticists to shrug their shoulders and claim that it is not reasonable to expect a solution to the problem of freedom of will through the development of robots."[89] Nevertheless, there is a distinct difference of focus between the philosopher and the roboticist's approach. The philosopher's conception of autonomy is related to issues involving the source or origin of an agent's pursuits (namely, that they issue from the agent itself), whereas the roboticist's conception of automaticity deals only with degree of human

[87] Haselager (2005) p 529
[88] Haselager (2005) p 519
[89] Haselager (2005) p 522

intervention during a particular performance. Haselager argues that this difference of focus "provide little ground for a debate between robotics and philosophy"[90] over the issue of autonomy. However, this does not render the two views entirely incompatible. A full account of autonomy ought to respect both an agent's freely chosen goals and the ability to independently pursue those goals without the intervention or assistance of other agents[91].

Haselager identifies a third conception of autonomy that is implicit in the philosopher's stronger view but which is also relevant to the roboticist's project. An autonomous agent doesn't merely have its goals, but it also *owns* those goals in the sense that the identity and integrity of the agent is partly constituted by the goals it freely chooses to pursue. Haselager claims that this issue of ownership is relevant both to philosophy and robotics because ownership is at least partially explained by a system's ability to organize and maintain it*self* as a purposive whole.

> **Ownership: "**Fundamentally, what makes my goals mine, is that I myself am at stake in relation to my success or failure in achieving them... Autonomy is grounded in the formation of action patterns that result in the self-maintenance of the embodied system and it develops during the embodied interaction of a system with its environment."[92]

In other words, the autonomy of a system depends on its ability to maintain a dynamic but persistent equilibrium. Maturana and Varela (1987) call such systems 'autopoietic': "An autopoietic system is a homeostatic machine, and the fundamental variable it aims to maintain constant is its own organization."[93] This ability is a characteristic feature of living systems.[94] Roboticists developing automated systems have a deep interest in such systems, since it is precisely the machine's ability to

---

[90] Haselager (2005) p 522

[91] Indeed, the extent to which an agent's independent pursuit of goals is 'automated' in the conventional sense of 'unconscious' or 'unthinking' appears ripe for philosophical analysis. I think that Haselager is too quick to dismiss the roboticist's definition as lacking philosophical import, but these issues lie beyond the scope of this chapter.

[92] Haselager (2005) p 523

[93] Haselager (2005) p 527. See also Susi & Ziemke (2001), and Keller (2008).

[94] See Emmeche (2007): "… organismic embodiment is the first genuine form of embodiment in which a system becomes an autonomous agent "acting on its own behalf " (cf. Kauffman 2000), i.e., taking action to secure access to available resources necessary for continued living."

maintain its integrity during a performance allows its human operator to relinquish control over its operations. On Haselager's view, it is the ability for self-maintenance that makes an agent's goals *actually matter to it.* "Importantly, homeostasis involves more than just keeping a variable constant through the use of feedback, as in the case of a thermostat regulating temperature, in that the homeostatic system necessarily depends, for its own existence, on the self-regulation. A malfunctioning or incorrectly set thermostat need not suffer from the negative consequences it produces, but a truly homeostatic system always will."[95]

However, one might be skeptical that anything could ever matter to a machine, even when it can self-regulate in the roboticist's sense. A close cousin of the ownership view is Haugeland's (1992) discussion of commitment. Haugeland claims that a machine's goal-driven behavior is at best *ersatz* behavior, since the machine could not in principle commit to the goals to which it is set. "The point is not that the standards are *given* to the robot; rather, the robot doesn't 'have' them at all—they remain entirely external to it."[96] Haselager admits that current machines are not embodied in the right way to be considered genuinely autopoietic, but considers this a legitimate avenue of robotics research. The challenge Haugeland raises against Haselager's view of ownership and commitment in machines is interesting, but it will raise issues related to the philosophy of mind that Lovelace's objection explicitly seeks to avoid, and will therefore take us far afield of the debate over autonomy in the relevant sense of 'independence' at stake between the roboticist and the philosopher. With these 3 conceptions of autonomy introduced, I now return to that debate.

The emphasis on 'originality' in Lovelace's formulation of machine autonomy skepticism appears to align itself with the philosophical conception of autonomy in Haselager's second definition; since the machines are not the originators of their performances, they appear to lack genuine autonomy. However, Lovelace's skepticism is not limited to concerns over the origins of the *goals* to which the

[95] Haselager (2005) p 526ff
[96] Haugeland (1992) p 302

machine is directed, but is instead targets the sources of the machine's performative abilities generally. Selecting goals may be among those performances, but does not exhaust the influence of human design and control on the behavior of the machine. Thus, Lovelace can admit to cases where a machine genuinely self-selects its performances, and therefore satisfies Haselager's second definition of autonomy. Even still, Lovelace can maintain that the machine's goal-selecting performances still originate from human design: it was programmed to engage in that goal selecting behavior, and therefore those performances are not properly attributed to the machine. The form of this argument should be familiar from Lovelace's concession in the classic debate, and should therefore be understood as a direct result of the DNTA that grounds her position.

This result is more surprising than it might first appear, so it will be helpful to see how the argument works in practice. Consider the use of evolutionary robotics[97], which trains robots to accomplish some task through selective pressures enforced across generations controlling software that change with variation according to algorithms inspired by biological evolution. This process purports to circumvent the apparent need for *offline* specification of goals by allowing a version of 'natural selection' to shape the system, such that the machines 'naturally' acquire goal-directed functions without the programmer's explicit intervention prior to the machine's performance. Thus, Haselager claims that "evolutionary robotics is relevant to the topic of autonomy since there is less need for the programmer and/or designer to 'pull the strings' and shackle the autonomy of the evolving creatures, because the development of robots is left to the dynamics of (artificial) evolution."[98] After several generations, a properly evolved robot will have acquired functions and behaviors that were not programmed in advance, but issue from the machine's own selection history. Haselager notes that although this method ought to work in theory, by analogy to the natural development of biological organisms, in practice roboticists are often tasked with carefully constructing the environment and the

[97] See Meyer et al (1998)
[98] Haselager (2005) p 524. See also Nolfi (1997) and Nolfi (1998).

development of the robots in a way that undermines the clear autonomy of the machine. Haselager quotes Nolfi (1997]: "Often, then, 'some additional intervention is needed to canalize the evolutionary process into the right direction' (Nolfi 1997: 196), keeping the designers well in control of their robots, even if the strings to which the robots are tied may be less visible."[99]

The relevant conception of autonomy at work in Nolfi's criticism is clearly the stronger philosophical conception of Haselager's second definition, since it concerns the offline evolutionary programming and environmental construction employed *prior* to the robot's online performance. From the roboticist's perspective, none of this offline tinkering should matter to the purported 'autonomy' (automaticity) of the machine. Thus, Lovelace can accept that the robot's eventual *online* performance is a genuinely self-issued goal-selecting performance, therefore meeting the criteria of Haselager's second definition, by taking a distinctly roboticist's perspective on the machine's behavior considered only in terms of its online performance. This interpretation would accord with Lovelace's friendly concession to the proponents of artificial intelligence discussed in the previous section. Lovelace's skeptical objection to machine autonomy is not that the offline tinkering somehow renders the machine's online performance illegitimate as an autonomous action. Lovelace's objection is more simply that the machine is only capable of such a performance due to "what we know how to order it to perform". It is self-issuing goals *because that's what we built it to do* during our offline construction. In other words, Lovelace's objection is not the result of a strong philosophical conception of autonomy, but instead targets and rejects the distinction between "online" and "offline" behavior that motivates the roboticist's conception of automaticity. Lovelace is effectively arguing that an explanation of the machine's performance must take into account the offline ordering, so the roboticist's attempt to circumscribe some subset of performances as "online" and therefore independent of prior construction is illegitimate. Whereas an AI skeptic like Haugeland might accept the online/offline distinction and

[99] Haselager (2005) p 525

argue that while the machine's construction offline is irrelevant, its performance online is nevertheless ersatz and not autonomous, Lovelace's objection takes the distinct form of conceding that the machine's online performance may be genuine, but nevertheless the offline construction *is* relevant, and it is on the basis of this offline activity that the machine is not autonomous. In this way, Lovelace's objection is clearly the result of a strong version of DNTA which stubbornly insists that as a technical artifact the machine's human origin and design is always in play even in cases of genuinely self-issued performances on the machine's behalf.

This shows Lovelace's argument to be stronger than most commonly accepted versions of DNTA, which typically limit the mind-dependence of artifacts to only their teleological aspects.[100] Call weak DNTA the view that holds that machines inherit *only* their goal-oriented purposes or functions from human minds. On this view, machines would inevitably fail to meet Haselager's second definition by their very natures. It is this weak form of DNTA that generates the apparent incoherence of the very notion of machine autonomy: only *we* can set the goals, purposes, and functions of a machine. The weak dual natures theorist will nevertheless hold that artifacts do enjoy *some* mind-independence simply in virtue of being physical objects obeying mind-independent physical laws, but this sense of independence is trivial and does no philosophical work, since those physical processes can bear no functional, intentional, or teleological goals independent of the activity of some mind. The weak dual natures theorist might be squeamish about using the word 'autonomy' to describe whatever physical (non-functional) independence machines have, but presumably they would have no problems accepting the roboticist's account of automaticity as a useful guide for understanding the limited role its human

[100] See footnote 4 above, especially Kroes and Meijers (2004) and McLaughlin (2001). One notable exception to this view is Thomasson (2007), who holds a version of DNTA that is not limited to intended functions. "The creator's intentions generally (whether or not they specify an intended *function*) are most relevant to determining whether or not her product is in the extension of an artifactual kind term." (p 8) Thomasson's view is discussed in detail in section 4.

counterparts play in a machine's online behavior. I take a view like this to represent the consensus position with respect to machine autonomy.

Lovelace appears to hold the converse view. If self-issued goal-selecting performances are sufficient for being autonomous in the philosopher's sense, Lovelace holds there is no incoherence in attributing such performances to the machine. Lovelace's strong dual natures view holds that the total behavior of the machine, and not merely its *function*, is the result of our explicit design; although it carries out its performances according to mind-independent physical laws, the performances are essentially influenced by our design. Our influence over the machine is present even when it pursues goals that are genuinely self-issued, and even when it self-regulates in order to maintain homeostasis. Thus, Lovelace's argument is ultimately an objection to the idea that machines enjoy *any* mind-independence. Lovelace is targeting the roboticist's assumption that we can profitably distinguish between online and offline behavior when attempting to understand the influence and control a human has over the machine. On Lovelace's view, we always have influence over the machine; our influence is not limited to setting up the initial 'offline' conditions and letting the automated physical system take care of the rest, but is continually at play even during the machine's 'online' performance. The physical attributes of the machines performances are also those attributes which our orderings selected it to have. While certainly compatible with the DNTA view that artifacts bear a relation to minds "in their very natures," Lovelace's position here nevertheless appears stronger than the weak view discussed above.

Although I have argued that the two views are distinct, the weaker view bears many of the crucial features of machine autonomy skepticism. In particular, it is committed to the position that so-called 'autonomous machines' pose no *special* legal, ethical, or metaphysical challenge that isn't accounted for by our normal theories of technical artifacts as mind-dependent tools. However, the target of this chapter is Lovelace's argument, and specifically her stronger dual natures account. In the

following sections I will argue that Turing's response to Lovelace is sufficient for undermining her strong dual natures theory; if my interpretation of Turing's argument is successful, it will also show the weaker view to be inadequate. Interestingly, this analysis implies that not only are weak dual natures theorists committed to the denial of the possibility of artificial intelligence, but that they also have some stake in the philosophical debate over autonomy more generally. The fact that dual natures theorists are committed to a position in the debate over autonomy shouldn't be surprising, since their characterization of technical artifacts is only substantive on the grounds that human minds are not subject to the limitations and causal origins of machines.[101] Philosophers who wish to avoid such a view will be urged to abandon the dual natures account, or at least to acknowledge that the mind-dependence of artifacts is not a one-way street, but is symptomatic of a deeper *inter*dependence between humanity and technology. This is precisely the route taken by Clark (1999) and other interactionists discussed in section 6 of chapter 1.[102]

3.5 Epistemic autonomy

Turing's sustained treatment of the Lovelace objection in his 1950's paper is partially explained by his sympathies with the motivations behind Lovelace's appeal to DNTA. Since Lovelace concedes that machines are not limited in their potential performances qua machines, Lovelace has no objection to the possibility of machines convincingly passing the imitation game or any of the more creative and robust

[101] This is precisely what is at work in Haugeland's (1992) distinction between genuine intentionality and the 'ersatz' intentionality of machines, and is also at work in Searle's (1990) distinction between 'derived' and 'original' intentionality.

[102] See also Clark (2008) (discussed at length in Chapter 2), Haraway (1984) and (1991), Flanagan and Owen (2002), Emmeche (2007), and many others. For instance, Emmeche says of this anti-DNTA literature, "We will only use these sources to emphasize that the machine is not any "alien" entity. Though usually made of inorganic materials, any machine is a product of human work and its societal network of skills, ideas, thought, institutions, etc., and thus, we reify or blackbox the machine when we see it as a nonhuman entity... Marx was probably the first historian to make this point which is also a point about embodiment. On the level of human embodiment, tools, technical artifacts, machines, and thus robots, all embody human co-operative division of labor, and of course, in our society, knowledge-intensive machines embody societal forms specific for a "post-industrial," "late capitalist", "risk" or "knowledge" society."" (p 475)

variations of the so-called "Turing Test" that have been proposed since Turing's original paper.[103] Thus, Lovelace appears to grant Turing the very position he sought to prove with his thought experiment, that machines are capable of performances on par with human beings. Furthermore, Turing himself is not interested in providing a working definition of thinking, and claims that the question "can machines think" is itself "too meaningless to deserve discussion."[104] As argued in Chapter 1, the comparison to human minds inherent to the indistinguishability requirement of his imitation game is best understood not as Turing's own hypothesis on the behavioral requirements for thinking, but rather as a thought experiment intended to be intuitively compelling to even the harshest critics of 'thinking machines'. Turing's test is widely recognized as a sufficient, but not necessary, condition for intelligence,[105] and this feature of the test suggests that Turing himself believes that there will be machines widely regarded as intelligent but fail to meet the indistinguishability requirement. Thus, Turing is also sympathetic to Lovelace's downplaying of the relevant analogies to human minds characteristic of the original debate, despite the fact that his own thought experiment encourages such comparisons as a concession to his critics.

The Lovelace objection poses a serious threat to Turing's goal of 'fair play for machines', since it appears to be grounded in facts about the construction of technical artifacts, and not in the unfounded biases against machines he worked so hard to dispel. Turing provides two responses to Lovelace in his original 1950 paper. First, Turing points out that sometimes machines surprise us, or otherwise act in ways that we do not (as a matter of fact) anticipate. Most of the computations a computer generates are not performed independently by its user or designer in order to verify the results; therefore, the computer will often generate results its user had not expected. Turing does not expect this response to convince the proponents of the Lovelace objection. However, he returns to the Lovelace objection in the

[103] See Saygin et al (2000)
[104] Turing (1950)
[105] See, for instance, Block (1981)

final section of the paper, and suggests the strategy of building machines that learn. In light of the analysis from the previous sections, it is clear that Turing's discussion of learning machines is aimed at undermining the apparent mind-dependence of machines central to the dual natures theory. This aspect of Turing's discussion will be picked up in the next section.

However, the standard interpretation of Turing's responses focuses on two features of the Lovelace objection. The first involves the idea of "originality" that Lovelace identifies as missing from her machines, and focuses the discussion of creativity as a form of mental activity and the role it plays in the design of artifacts, and whether this creativity can be found in our machines. This discussion takes on the familiar arguments from the classical mirroring debate, and will be set aside here. The other, more prominent interpretation of Lovelace's objection (sometimes combined with the first) treats the debate as an epistemological issue, and not one of autonomy or the mind-dependence of technical artifacts. The emphasis in Lovelace's objection is placed on *what we know* about the machine; on this interpretation, the machine does not act independently of what we know of it. Turing's advocacy for learning machines in response is usually interpreted as a defense of the machine's epistemic opaqueness: that at least in the case of learning machines, there is something about the machine that we do not, and in fact *cannot*, know. For instance, Abramson (2008) gives a sustained treatment of Turing's discussion of the Lovelace objection:

> "Turing's focus is on the epistemic relationship between the creator of the machine and the machine created. Turing clearly is committed to the view that in order for the actions of a machine to be truly its own, and not the achievements of its creator, there must be an epistemic–limitation on the creator with respect to the machine's behavior."[106]

While Abramson's interpretation agrees with mind insofar as it identifies the critical issue of the Lovelace objection to be the relationship between the designer and the machine, I will argue in this

[106] Abramson (2008) p 160ff

section that the epistemological reading of this relationship misses the force of Lovelace's strong objection, and consequently underappreciates Turing's response.

To be fair, the robotics literature has long recognized the importance of a kind of epistemic autonomy on the part of the machine. Epistemic autonomy is crucial to the successful automation of systems that must respond appropriately to their environment; in particular, it is important for the machine to have its own 'sense' of the appropriateness of its responses. Prem (1997) defines *epistemic autonomy* in the sense relevant to the discussion of robotics as follows:

> "The deeper reason for this strategy lies in the necessity to equip embodied systems with *epistemic autonomy*. A robot must be able to find out whether its sensor readings are distorted and, even more importantly, exactly when a measurement has occurred. The correct *interpretation* of sensor readings is of vital importance for the generation of useful system actions. This epistemic condition on autonomous systems is central to all ALife models, indeed to all living systems. It arises in autonomous embodied models, because no humans are available to interpret the data and a pre-selection of valid data is impossible for practical reasons. (This is in sharp distinction from purely formal domains, in which the interpretation of simulated sensors can always be reduced to formal constraints. The ability for whiskers to break, however, is usually not modeled in any robot simulation.)"[107]

Given our discussion in section 3, the relevant notion of autonomy in this passage is clearly the roboticist's, since it concerns the system's online behaviors that fall outside the oversight of its human designers. The practical limitation on human oversight identified by Prem is actually a stronger version of Turing's own discussion of surprise in his first response to Lovelace:

> "Machines take me by surprise with great frequency. This is largely because I do not do sufficient calculation to decide what to expect them to do, or rather because, although I do a calculation, I do it in a hurried, slipshod fashion, taking risks. Perhaps I say to myself, "I suppose the Voltage here ought to be the same as there: anyway let's assume it is." Naturally I am often wrong, and the result is a surprise for me for by the time the experiment is done these assumptions have been forgotten. These admissions lay me open to lectures on the subject of my vicious ways, but do not throw any doubt on my credibility when I testify to the surprises I experience."[108]

[107] Prem, E. (1997)
[108] Turing (1950)

Turing reasons that while he should in principle be able to anticipate the machine's behavior, in practice this is often not the case, usually on account of his own errors. However, Turing is discussing digital computers as formal systems and not embodied robots, which makes it plausible to think that, had he done his homework properly, he could have anticipated all the behaviors of the machine. Prem points out that when dealing with real-world machines we can't always anticipate its behaviors even in principle since the physical system is subject to causal interactions that we may not have explicitly modeled, like breaking whiskers. Turing's argument comes close to acknowledging Prem's embodied/formal distinction, since his argument deals with the electronic hardware underlying the performance of the computer and not with its purely formal structure.

The critical point in Prem's definition is that epistemic autonomy depends on the system's epistemic *authority* with respect to its inputs, and the types of performances this authority engenders. A machine is epistemically autonomous if it takes itself (and no other operator or designer) as the relevant authority in its judgments. As I argued in the previous section, Lovelace would have no problem granting the robot this kind of epistemic autonomy since it only specifies a certain kind of automated online performance, and on Lovelace's view machines are not limited in their possible online performances. The fact that there are no humans around to monitor and interpret these inputs is not a necessary feature of this autonomy but is entirely coincidental; a machine can take itself as an authority regardless of what others know about its internal states. If a human were available and aware of the calculations the systems makes online, the system would still be autonomous in the relevant sense if it took itself (and not its operators and overseers) to be the relevant authority. If some aspect of its operation is nevertheless opaque, this represents a merely *practical* limitation on our knowledge of the machine, not a condition on its autonomy. It is *because* of this practical limitation (we don't always know what's going on inside the machine) that the machine is designed to be epistemically autonomous (so that it can take

care of itself independent of what we know), but that autonomy doesn't depend on its epistemic opacity.

Against this objection, Abramson argues that Turing is committed to a *necessary*, and not merely practical, limitation on what we can know in advance about an intelligent machine's performances, by appeal to a theory on the essentially creative nature of intelligence. Abramson outlines the argument as follows:

> "1. Creativity involves the ability to originate at least something.
> 2. Following a set of rules intended by the creator of the rules to bring about a particular behavior does not involve originating anything.
> 3. Therefore, machines that are programmed to have intended behavior are not creative.
> 4. Creativity is essential for intelligence.
> 5. Therefore, intelligent machines must have behaviors not intended by their creators.
> 6. Therefore, intelligent machines must not be programmed as conventional computers are."[109]

Abramson interprets Turing's suggestion to build learning machines as a method for developing unconventional computers that behave in ways its designers do not intend. He justifies this interpretation from a variety of interviews and lectures Turing gave in the early 1950s, most prominently a radio interview in 1952:

> "From this point of view one might be tempted to define thinking as consisting of 'those mental processes that we don't understand'. If this is right, then to make a thinking machine is to make one which does interesting things without our really understanding quite how it is done."[110]

While Abramson admits that Turing is speaking with tongue in cheek, he takes Turing's suggestion seriously and interprets Turing as demanding that an intelligent machine must not simply be epistemically autonomous but its designers *must not know what it will do in advance*. In other words, Abramson argues that intelligent machines must be epistemically opaque. Drawing from this quote and discussion above, we can specify three ways in which a machine might be epistemically opaque:

[109] Abramson (2008) p 163
[110] Turing (1952), as quoted in Abramson (2008)

1. **Weak opacity**: The machine might do something that its designers do not (as a matter of fact) anticipate or predict. In other words, the machine might surprise us. This is not to say that we couldn't predict the behavior in principle, but just as a matter of course we have not worked out its behavior in advance.
2. **Strong opacity**: The machine might do something that its designers *cannot* predict. There are some mysterious or unexplainable phenomena that cause the machine to behave in a way that its designers could not have in principle anticipated.
3. **Mechanical failure**: the machine does something as the result of some error, either in its programming or physical constitution, that was not the intended result of its original design.

Neither the first nor the third of these types of epistemic opaqueness will be sufficient for satisfying Abramson's argument. In the first case, even if I do not as a matter of fact anticipate the machine's performance, it is still performing as I ordered it and therefore not originating anything new. In the case of mechanical failure, although such failure is not part of the machine's explicit design that doesn't imply that it is a creative act, and so is insufficient for establishing that the machine itself has created something original. Thus, Abramson is committed to the second type of opacity as a necessary condition on intelligent machines.

Abramson is not alone in associating Turing with this strong opacity requirement. Bringsjord (2001) formulates what he called the "Lovelace Test" as a test for intelligence based on this requirement. On Brinsjord's view, a machine passes the Lovelace test when the following conditions are met:

> "1. M outputs o;
> 2. M's outputting o is not the result of a fluke hardware error, but rather the result of processes M can repeat;

3. H (or someone who knows what H knows, and has H's resources) cannot explain how A produced o by appeal to Ms architecture, knowledge-base, and core functions"[111]

That H "cannot explain" how A produced o cannot be a matter of coincidence or laziness that might "surprise" us in Turing's sense, but must be due to the principled inexplicability of the machine's performance qua performance. The challenge for such views is to specify how a machine that we build is capable of performing in these mysteriously opaque ways. Both Abramson and Bringsjord appeal to the halting problem as a way of demonstrating that computers can, in fact, behave in unpredictable ways, but it is far from clear that the halting problem is itself sufficient for responding to this interpretation of the Lovelace objection. The fact that there is no general solution to the halting problem implies that there is no general shortcut for determining if a given Turing machine halts on some input save actually running the machine and seeing if it halts. When a given machine halts it might surprise us, but the fact that there is no solution to the halting problem doesn't entail that a machine is epistemically opaque or that its behavior is inexplicable. At best it shows a practical limitation in anticipating the machine's behavior, but we might still in principle know precisely which operations it will perform given its state an input.

Complications with the halting problem aside, there are a number of reasons to be concerned with this epistemological interpretation of both the Lovelace objection and Turing's response. For one thing, there seems to be little textual support for the interpretation. Turing's half-joking appeal to the inexplicability of intelligent behavior is unconvincing at best, and a serious misunderstanding at worst. On my interpretation, which seems a more natural and charitable reading, Turing's appeal to inexplicability has the same rhetorical force as his appeal to consciousness or souls,[112] namely that insofar as this property is assumed of human performances, we might use the same criteria to apply that judgment to machines. In none of these cases should we take seriously the proposal that these are

[111] Bringsjord (2001), as paraphrased by Abramson (2008) p 164
[112] See the discussion of these objections in Chapter 1.

necessary features of thinking; Turing's entire strategy undermines any necessary requirements for intelligence whatsoever. My reading of Lovelace's objection also finds straightforward textual support in tension with Abramson's interpretation. When Turing begins the final section of the paper he explicitly returns to the Lovelace objection and paraphrases as follows: "Let us return for a moment to Lady Lovelace's objection, which stated that the machine can only do what we tell it to do."[113] While this reformulation of the objection is consistent with my interpretation, it completely abandons any pretension to an epistemological limitation condition as Abramson understands it.[114]

Beyond textual interpretation, Abramson's reading straddles Turing with a *necessary* condition for intelligent behavior, which he explicitly admits is contrary to Turing's test as a merely *sufficient* condition for intelligence. Abramson attempts to circumvent this worry by arguing that any machine that meets the sufficiency condition will as a matter of fact (though not of *logical* necessity) also meet the epistemic limitation condition. "On the reading I am presenting, Turing makes the empirical claim that, as we try to build machines that pass the Turing Test, we will have poor success with machines that contain surveyable procedures for anticipating questions. This, in fact, has been the case."[115] Abramson appeals to the Loebner Prize transcripts as examples of machines that fail to convincingly pass the Turing Test. However, he does not show that the reasons for these failures are due to the epistemological *clarity* of these systems, only that the systems that do well tend to be opaque. It is also

---

[113] Turing (1950)

[114] Abramson, in personal correspondence, responds to this interpretive point as follows:

> "The reason that Turing offers this formulation is because it is true, whereas the formulation that Lady Lovelace offers is false. The section that follows shows some ways to build machines that, although they satisfy Turing's formulation, do not satisfy Lady Lovelace's.
>
> Simply put, computers really do follow a course of action which is entirely determined by their initial state and their input. It does not follow from this that they only do what we have knowledge that they will do, as determined by our knowledge of what their initial state and input are. That we *can't* know this, in general, follows from the unsolvability of the halting problem."

However, Turing doesn't indicate that he is changing the goalposts, as Abramson suggests, and if the goalposts have changed its not clear why he takes the Lovelace objection as seriously as he does. Notice also that Abramson is implicitly committed to a weak version of DNTA.

[115] Abramson (2008) p 163

not clear why surveyability is a criterion for explicability; I might not be able to track the stepwise computations of a program running at speed, but this seems far short of the inexplicability required to generate the conclusions Abramson draws.

Abramson's interpretation also saddles Lovelace with a similarly implausible view that seems hard to take as seriously as Turing does: namely, that we can know everything about our artifacts. A closely related position might be Thomasson (2007) argument that "makers of artifacts are (as such) guaranteed certain forms of immunity from massive error about the objects of their creation."[116] Thomasson argues that artifactual kinds are essentially the products of transparent designer intentions, and thus their designers are granted a kind of epistemic privilege over their natures. While there are relations between Thomasson's and Lovelace's views, since they are both derived from a dual natures account[117], they differ in their epistemological import. Thomasson is not committed to complete knowledge of the *performances* of artifacts, in the sense of predicting what a machine will do in a given circumstance. The maker's epistemological privilege is limited to knowledge of the abstract kind of which the artifact is an instance. Similarly, Lovelace doesn't appear to be committed to the view that we have complete *knowledge* of the system's performance; at most, in her original formulation, we have knowledge of only what we have ordered it to do, or more precisely we have knowledge of *how* to order it to perform. Lovelace isn't arguing for an epistemological privilege over the behaviors of the machine, and she's certainly not committed to the position that we'd know if a Turing machine will halt. Instead, she's arguing for a sustained human influence over the machine's 'online' performances.

3.6 Conclusion: learning machines and machine participation

[116] Thomasson (2007) p 16. See also Thomasson (2008)

[117] Thomasson (2008) argues that artifactual kinds (unlike other natural kinds) must have mind-dependent natures; she claims this is part of the "very idea of an artifactual kind". While I don't object to the proposition that at least some (and maybe most) artifacts have such natures, nor do I object to Thomasson's common-sense ontology, my arguments in this dissertation implies a rejection of Thomasson's view insofar as machine participants are artifactual kinds that do not depend on minds for their natures.

Finally, Abramson's interpretation seems to underappreciate the importance of Turing's suggestion for building learning machines. On his interpretation, Turing offers the suggestion in response to the Lovelace objection presumably on the grounds learning would produce machines that are epistemically opaque. But this is an awkward reading of the proposal. The process of learning is not simply a method for imposing epistemological distance between a teacher and pupil. Consider the relationship between a (human) instructor and teaching her (human) student how to do perform the algorithm for long division. At first, the student will do poorly, and require the repeated assistance of the instructor. If the instructor is asked early in the learning process, "did the student complete the problem on his own?" the instructor would be forced to answer in the negative: the student is only capable of completing the assignment with her assistance. However, after considerable practice and patience the student will eventually grasp the method, and the instructor will be able to answer the same question in the affirmative. This is a case of genuine learning of the familiar and entirely uncontroversial sort; perhaps different students will have different needs in the course of the learning process, but other things equal the learning process results in competent and independent performance. In the course of learning the algorithm, the student develops the capacity to perform the algorithm reliably and on her own authority, without recourse to any other agent. The end result of learning, beyond the skills and knowledge acquired, is the epistemic autonomy to employ those skills as required. The instructor can positively identify that the student has learned long division on the basis of their autonomous performance in long division problems, while knowing full well what steps will be employed, in what order, as the student executes the algorithm. The student need not surprise the instructor, and the instructor might be fully capable of explaining how and why the student is able to successfully solve a long division problem by appeal to the methods that were instructed. What makes this a case of learning is precisely when the student is able to perform the activity independently, without directed guidance and instruction from his teacher. Having learned, the student is judged to be

a competent performer of that task in terms of the performance itself and the standards by which such performances are normally judged. In the case of the human learner, the "offline" activity conducted during the learning process may very well have some impact on the nature of the eventual performance. But in the case of humans, we don't typically judge the instructional activity between teacher and student as relevant in judging some later performance as a competent agent. Once I've learned to do long division, for instance, my eventual performances of the algorithm later don't (necessarily) bear the mark of that original instruction event.

To sum up, in the case of human learning, the process does not result in epistemological opaqueness between student and teacher, and the student doesn't typically "bear the mark" of the instructor in their independent competent performances of what it has learned.[118] Instead, the process of learning results in the student's eventual epistemic authority and autonomy, and acquires an agency independent of the instructor's lesson. If learning is not explained in terms of epistemological opacity in the normal case, then Abramson's interpretation fails to do justice to the Lovelace objection, and subsequently to Turing's response. Lovelace's objection is not worried that our knowledge of artifacts prevents the possibility of their creativity, and Turing's response is not arguing in favor of creating inexplicable machines. Instead, Lovelace is arguing from a conception of artifacts that appears to prevent the possibility of machine autonomy to the entire category, and Turing's response is an attempt to demonstrate how we might come to treat even these technical artifacts as genuinely autonomous. If our machines learn to acquire their behaviors over time and adjust their performances on the basis of their experiences of the world, then the resulting performances are not merely the result of what we've ordered the machine to do, and therefore the machine has taken on a character that reflects neither the machine's use as an instrument nor its design as an intentional artifact. Such an artifact would not merely be a tool, but would be an independent agent of a different sort entirely.

[118] Some teacher-student relationships might leave a special mark on a student's performance, of course, but this is hardly necessary or universal in instances of learning.

In Chapter 1 I introduced this independent artifactual agent as a "machine participant", and I conclude that this artifact satisfies the goal of our discussion in these two chapters of finding machines that are not tools. A machine participant is an artifactual agent whose functional character is importantly mind-independent in both the sense of use and design. This is not to say that aspects of a machine participant's nature won't continue to be use- or design-dependent; in other words, they are still artifacts. But their natures are not exhausted by their artifactual or instrumental character, and must be considered as agents in their own right. This agency is also not exhaustively explained by the physical characteristics alone, insofar as these characteristics fail to account for the functional role performed by the machine as a participant. In this way, machine participants are unique among the artifacts, and should be given special philosophical consideration in virtue of this autonomy. This shows the dual natures account of artifacts is inadequate for treating machine participants, and cannot constitute an exhaustive account of artifacts and the role they play in human organizations. This argument does not depend on machine participants as being more "mind-like" than other artifacts, since we can continue to treat very mind like machines[119] as tools and ignore their agency, and conversely, some machines radically different from minds (like Google) might nevertheless qualify as participants in virtue of their behavior.

This project continues in Chapter 3, where I will explicitly consider technological organizations of this type, assuming a framework of machine participation.

[119] Including, in the limit case, treating other human beings merely as tools, a custom that is unfortunately all too common in human history.

Chapter 4

The Organization of Technological Networks

4.1 Abstract

Network theory provides a framework for treating the complex, dynamic relationships between diverse, independent agents within an organized system. I give an extended treatment of Levy and Bechtel's (2012) discussion of organization and its role in mechanistic scientific explanation, and discuss the importance of organizational degree in graph theoretic terms. This discussion provides the background for a network theoretic treatment of sociotechnical systems that can serve as an alternative to the mind-centered models of technology discussed in chapter 1. Using a variety of case studies, including the flow of traffic and ultra-high frequency trades, I propose that the organizational dynamics of these networks are best understood as emergent phenomena, where system-level properties might arise distinct from the properties of their components. I argue that this approach allows for a decentralized theory of autonomy that does not depend on mental activity and that satisfies Turing's call for fairness. With this theory in hand, I identify a few key technological networks, including the internet and the much larger attention economy in which it develops. In some instances, these networks can demonstrate emergent behavior sufficient for rejecting the standard view of artifacts.

4.2 Some examples of machine participants in complex networks

In chapters 1 and 2 I discussed use and design as primary relationships between minds and tools. By reflecting on Turing's discussion of learning machines as a response to the Lovelace objection, I concluded that some machines demand explanations as participants, and that this account of machine participants is not available to conceptions of machines that treat them necessarily as tools. Machine participants are artifactual agents with functional characters developed in the course of their engagement with an agential community of interdependent users and tools. Since this functional

character depends on the machine's history of community engagement, that character cannot be understood in mind-dependent terms of use and design typically used to treat tools. Therefore, machine participants provide some reason for rejecting the "dual natures" account of artifacts, or at least thinking it incomplete. Machine participants might also fail to appropriately mirror the mental capacities of the creatures we take to have minds, and therefore cannot be understood in the terms typically used in the classic artificial intelligence debate. If these accounts fail to adequately address the activity of machine participants, then we require an explanatory framework of artifacts that does not center on minds or reduce functional capacities to extensions of the activity of minds.

But before going farther, perhaps it will be useful to provide more detailed examples of machine participants. In this section, I will provide three examples of machine participants of the sort I have in mind. With these examples in hand I'll sketch a network-theoretic framework for thinking about technological systems, where machines play roles as components with many interdependent relations to other (potentially human) components. In contrast to the dual natures account of artifacts, my framework is nonreductive and allows for explanation of machine participants to emerge naturally from an analysis of the organization of the network.

In section 3, I turn again to theoretical issues in a theory of organization in complex networks of the sort I find appropriate for treating technology. Following Levy and Bechtel's (2013) discussion of abstraction and the organization of mechanisms, I address the issue of organizational degrees and its relation to mechanistic explanation. Along the way I will introduce a variety of tools and concepts from network theory that are useful for thinking about mechanistic explanation generally and will be relevant for our discussion of technological networks in particular. I conclude the chapter with some discussion of the implications this explanatory framework has for the way we think about artifacts and the worlds we share.

I'll now introduce some examples of machine participants of the sort I've argued for at the level

of theory over the last two chapters. Our discussion of Turing suggests some obvious examples that I want to avoid for the sake of clarity. For instance, Cleverbot[120] is an online chat bot derived from a past winner of the Loebner prize, the official annual run of Turing's test for identifying intelligence machines. Although simple compared to human language production, Cleverbot is notable in that it acquires new patterns of speech on the basis of the conversations it has had with past interlocutors. In other words, Cleverbot is a learning machine engaged in exactly the sort of linguistic conversations recommended by Turing's imitation game. The behavior elicited by Cleverbot is not just the result of its initial programming or its prompted "use" by an interlocutor, but is also fundamentally characterized by speech patterns it has learned from others. Since Cleverbot has been engaged in conversations with people online now for many years, it has acquired all the charm and personality you'd expect for a chat bot raised online. On my view, Cleverbot is a machine participant, not because it is sufficiently mind-like or because it is programmed or used to be some particular way, but because its behavior has been developed through a history of activity as a participant in conversations, in the role of a conversational interlocutor. Cleverbot demands an explanation that appeals to this conversational history as a mind-independent fact about the artifact; this explanation is both functional/intentional in character, and is not merely the product of interpretive stances taken by particular human interlocutors (who may or may not treat it merely as a tool). Its participatory history a fact about the artifact necessary for explaining the behavior it elicits. Nevertheless, using an example like Cleverbot, whose participatory behavior depends on an appearance of intelligence, risks reviving all the old problems of the classic artificial intelligence debate, a stalemate that I've been keen to avoid. Instead, I will consider three examples meant to highlight the participatory status of the machine independent of any comparison with human minds.

[120] http://www.cleverbot.com/

The first is the case of lying robots.[121] Mitri et al (2009) describe an experiment in evolutionary robotics[122] where collections of social robots are selected across multiple generations for their ability to find food and avoid poison within a small field of activity. Their experiment is designed to give an evolutionary perspective on the development of communication signals within a social population. The robots are fitted with an array of lights and sensors they can use to signal to other robots their state in the field and the location of food or poison. The robots begin with random firings of blue lights, and in each generation robots are selected on how successfully they acquire food. This selection pressure towards finding food eventually manifests in an attraction to blue lights from other robots as an indicator of the food source, since there increasingly tend to be more robots (and hence blue lights) around food. However, this also results in overcrowding around the food; only a limited number of robots (smaller than the total population) can feed at any time. In this particular experiment, the robots that were most successful developed a disposition to *suppress* the firing of their blue light; in other words, the robots stopped signaling the successful discovery of food. In the context of a population that had developed the behavior of flashing blue lights to indicate food, these robots discovered that *misrepresenting* their success was the more successful strategy. Surprisingly, although this behavior for lying about food was developed within 36 generations, the general population was not able to compensate for this change in the information content of the signal for several dozen generations afterwards, thus demonstrating the impact that information suppression can have on the evolution of social signaling.

The lying robots are participants in the sense that the behavior (information suppression) is explained in terms of their selection history and their complex interactions with their fellow robots in

[121] Mitri, S., Floreano, D., & Keller, L. (2009).

[122] The picture of machine participation described in this dissertation has found natural applications throughout the field of robotics. Evolutionary robotics in particular has long struggled with the issue of divorcing human control and influence from the operation of the machine, and has developed theoretical tools and vocabulary that lend themselves naturally to the approach taken in this dissertation. See Meyer et al. (1998), Rohde et al (2010), and Chapter 2 of this dissertation for more discussion.

the shared task of finding food. The behavior of lying emerges from the interactions between these robots as they refine and make use of a common signal (the blue light) for indicating shared environmental resources. These machines are not "tools" in the sense appropriate for the use relation, in the sense that the information suppression behavior is being actively employed by some other user for the sake of some instrumental goal, and their behavior is only indirectly related to the experimental design. Although the evolutionary algorithms used for the experiment are inspired by an analogy to biology, there is no serious suggestion that these robots mirror important cognitive structures of humans or any other animal except insofar as both are to be explained in terms of their selection history. Therefore, the lying robots serve as an example of machine participants in terms of the framework I've introduced: they have developed and sustain a meaningful social practice whose content is not derived from the extended activity of any other mind. However, as discussed in Chapter 2, there may be some lingering worries that evolutionary robotics demonstrates insufficient autonomy from their experimental design, and perhaps this example will strike many as insufficient. Rather than rehearse the arguments from that chapter, I'll continue with more suggestive examples to give a feel for the position and its relevance. An example from outside the domain of robotics might serve us better.

So-called "ultrafast extreme events" (UEE) have become increasingly common in the world of high speed financial trading.[123] UEEs are financial transactions that take place in timeframes on the scale of milliseconds, far beyond the capacity for deliberate human response and control. These events cause dynamics in the market (like spikes and crashes) that cannot be explained in terms of human intentions precisely because of the speeds at which they take place; in fact, Johnson et al (2013) finds that the number of such events increases dramatically the more they fall below the threshold of human response. Understanding these events and their impact requires more than understanding the intentions and purposes of the businesses that use these events as tools to gain an advantage in the

[123] Johnson, N., et al (2013).

market, since the coarse-grained mental activity of use simply does not operate at the scales relevant for explaining these complex dynamics. Instead, Johnson et al. construct a model of many adaptive agents. Each agent in this model employs a variety of strategies to gain a competitive advantage, but only some of these agents are capable of performing at the speeds required for UEEs to occur. Their model predicts that agents will converge on a strategy that employs UEEs more frequently in order to optimize their competitive advantage. While the general strategy of "optimizing a competitive advantage" may be part of the business' strategic intentions in building and maintaining computers capable of high-speed trades, this intention is in general not capable of determining which specific pattern of trades are initiated. Instead, these trades are the result of evaluation and direct market engagement initiated by the machine itself as an agent of a complex, multi-agent, multi-scale system.

There are undoubtedly legal, political, and ethical challenges with the introduction of UEEs into existing human economic markets[124]. Although the view of machine participants will have implications for an analysis of those challenges, it is not the point of this dissertation to engage with, much less resolve, questions of those types. Instead, my concern throughout this dissertation has been to engage more basic questions concerning our methods for explaining and predicting the behavior of sociotechnical systems in which machine participants play an ineliminable role. Vespignani (2009) summarizes the position clearly in his short discussion of the challenges in predicting the behavior of sociotechnical networks; I quote the passage at length to highlight the convergence of the scientific challenges with the perspective I've attempted to articulate in this dissertation:

> "However, the biggest challenge in providing a holistic description of multiscale networks is the necessity of simultaneously dealing with multiple time and length scales. **The final system's dynamical behavior at any scale is the product of the events taking place on all scales.** The single agent spreading a disease or single node of the Internet that fails are apparently not affected by the multi-scale nature of the network, just as single molecules do not care about the multiscale nature of turbulent fluids. However, the collective dynamical behavior and our ability to conduct mathematical and/or computational analyses of techno-social systems are constrained by the multiscale

[124] For an introduction to this discussion compatible with the framework presented here, see Moor (2006).

> characteristic of the system. In the context of networks and techno-social systems, the multi-scale challenge is making its appearance now because of the availability of large-scale data sets. Thus, we have to develop appropriate formalisms and techniques, as researchers studying multi-scale physical systems (fluids, solids, distribution of masses in the universe, etc.) have done in the past. To achieve analytical understanding of techno-social systems and approach them computationally, we must find different strategies to deal with dynamical behavior and/or equations that work at very different characteristic scales but still influence each other. Such methods will finally allow the definition of layered computational approaches in which different modeling assumptions and granularities can be used consistently in the context of a general multiscale framework."[125]

I take the description of machine participants of the sort defended in this thesis to be a necessary part of any response to Vespignani's call for granular formalisms appropriate for describing such multi-scale complex dynamics.

Vespignani goes on to argue that generating these formalisms require 'network thinking', where individual agents are taken as nodes at different scales with different capacities for action[126]. Such networked models are commonplace in describing complex multi-agent systems, of which the financial markets are surely an important case. Today's markets are among the most complex sociotechnical systems[127] we've ever built, comprising many humans and machines in a variety of different capacities and complex dynamics both within and across organizational structures. Treating these as multiagent systems with diverse components allows us to explain their dynamics and subtle complexities at many scales, without assuming the relationships will be characterized in terms of metaphysical or instrumental relationships to typically attributed to intentional, minded agents at some fixed scale. Instead, multi-agent systems allow us to characterize the relationships between both human and machine components of complex systems in a variety of ways, some of which might involve functional relationships between machines that bear no relationships to mind. Moreover, the organizational dynamics of these complex networks might themselves give rise to important phenomena that demand

[125] Vespignani (2009)
[126] For an overview, especially as the concept appears in the artificial intelligence literature, see Ferber, J., & Gutknecht, O. (1998, July). See also Omicini et al. (2004).
[127] Kline, S.(1985)

explanation in terms of this network of relations.

In contrast, the dual natures account of artifacts is, in a sense, a reductive account of artifacts--not that artifacts are reduced to their physical constitution, but that the functional contributions of artifacts are reduced to the activity of the minds on which it depends for its very nature. Insofar as the functional character of machines is derived from the activity of human minds, the functional contribution of artifacts to a network can be represented entirely as extensions of the activity of those minded agents. Consequently, the relations among agents on the dual natures view takes the form of relations between agents of essentially homogeneous type, and an explanation of the organizational dynamics of a multifaceted sociotechnical network is lost. If the agents in our financial markets are all the extensions of some human mental activity, then the phenomenon of UEEs goes unexplained. If, on the other hand, the agents in the system are diverse and function outside direct human control, then you can build a model that explains the behavior.

On my view agents are the components of an organized network, and their agency consists in the relations between themselves and other components in the network. Natural systems are organized in myriad ways and at many levels of analysis, which raises issues about identifying the particular systems and components under question. This problem will be addressed in the next section. Before turning to a discussion of organization in complex networks directly, however, I want to introduce one further example of "machine participants" that will hopefully give the sense of the scope for which I mean the term. In Chapter 1 I briefly discussed the human-car cyborg assemblages in terms of agency and use. Scientists commonly use multiagent systems to model the behavior of traffic consisting of such agents,[128] and these models specifically allow for characterization of features like traffic flow and traffic jams as phenomena that emerge from these dynamics. Sugiyama et al (2008)[129] demonstrate the emergence of traffic jams in contexts without any design bottlenecks, among agents instructed to

[128] See Bellomo, N.,(2002); Paruchuri, P., (2002).
[129] Sugiyama, Y., (2008).

maintain a steady distance from the cars in front and behind them. Although these traffic jams emerge reliably, their emergence cannot be explained in terms of the intentions of any agents, either those on the road or those designing the experiment. Like the oscillating dances performed in the perceptual crossing experiments discussed in chapter 1, the emergence of a traffic jam is a product of quasi-autonomous participatory agents self-organizing in response to each other in a dynamic field of activity.

A traffic jam is clearly an artifact of human construction, and indeed does depend on human mental activity for is character. Human cognitive systems are simply poor at maintaining steady distances, and the slight variations that begin to emerge eventually cascade into variations large enough to experience as a jam. A population of well-programmed autonomous cars that perform better at maintaining a steady distance might avoid jamming, so the human components of the vehicle agents certainly matter to the dynamics of the traffic. But those dynamics can't be explained in terms of the design or instrumental intentions of any of those human agents. The jam is a genuinely emergent phenomenon, not explainable by reference to the properties of any component alone or even their simple linear sum, but must be explained by reference to the interdependent and nonlinear relations between components at many levels of organization. The traffic system is composed of vehicle agents, which themselves are composed partially of human cognitive systems. Understanding the relations between these components at various levels is necessary for understanding the flow of traffic.

And as a phenomenon, the traffic jam itself can be constituted as a participatory agent with whom I must contend: it can enrage me, it can prevent me from making an appointment, etc. These relations to the emergent phenomena of a traffic jam (which, again, emerges from a sociotechnical system itself composed of human agents) and still further organize our social lives, as when we plan around its existence or collectively plan for the development of our community infrastructure. Traffic systems are a paradigm example of participatory networks of organized agents of a variety of types, roles, and capacities, and appreciating the full complexities of these interactions across many levels of

analysis is necessary for describing these phenomena. The dual natures view is inadequate for treating our technological circumstance, and will be discarded in what follows.

The dual natures account of artifacts has received some criticism in the literature for being reductionist and insensitive to social dynamics that inform the nature and function of artifacts. Schyfter (2009) argues that social institutions play an important role for understanding the nature of artifacts. Vaesen (2010) argues that nonfunctional social facts about artifacts (like marketability, or easy of manufacturing) are also important for understanding the nature of artifacts and the role they play in human social systems. Heersmeek (2014) provides a recent critique of the dual natures account from an interactionist perspective very much like the one defended in this dissertation, where "cognitive artefacts are... neither defined by intrinsic properties of the artefact nor by the intentions of the designer, but by their function, which is established by the intentions of the user and by how it is used."[130] Heersmeek proceeds to describe a model of cognitive function "as an emergent property of the interaction between intentional, embodied agents, and cognitive artefacts." The resulting view is very similar to the one presented here.

However Heersmeek remains neutral on whether cognitive artifacts can themselves constitute embodied agents; in other words, it does not take a stand on the status of machine participants. While I'm sympathetic to these criticisms from the literature, these critiques are importantly distinct from the defense of participatory machines I've provided in these first three chapters. Moreover, my discussion connects the philosophical discussion closely with literature in artificial intelligence, robotics, cognitive science, complex adaptive systems, and network theory. This integrated approach is intended to bring together many important, existing discussions on the nature and future of technological development through what I believe to be an original contribution. I turn in the next section to discuss more abstract issues concerning organization in the context of complex networks.

[130] Heersmeek (2014)

4.3 Forms and degrees of organization

This section concerns Levy and Bechtel's (LB) recent attempt[131] to turn the philosophical discussion of mechanism in the sciences towards the issue of organization. Organization is an important feature in many of the systems we care about, but any precise characterization of the phenomenon seems to defy the reductive analytical techniques that have worked so well in other domains of science. The terms typically invoked in the discussion (emergence, complexity, function, entropy) cry out for the clarity of philosophical analysis, but in decades of research no single view has definitively claimed the consensus position[132]. There has been a resurgence of interest in mechanical explanations for their potential to fill the role of a systems-theoretic explanatory account, especially in the biological sciences. Richardson and Stephan elaborate:

> "Mechanistic explanation offers a different perspective on the understanding of the relationship of explanations at different levels or, perhaps better, the relationship of explanations incorporating different levels of organization. In mechanistic models, higher level explanations are neither redundant nor eliminable. There is genuine explanatory work done at the higher level, reflecting systemic properties. When this is so, and when that explanatory work is not captured at a lower level, the higher levels of explanation are not eliminable."[133]

LB employ the formal resources of graph theory to motivate their treatment of mechanistic explanation[134] in terms of organizational complexity, and link the discussion of organization to their research on mechanisms and their explanatory role within the sciences. In so doing, LB provide a much-needed introduction to conceptual resources of networks aimed at a general philosophical audience. However, the issue of organization is not the direct target of their paper. Instead, their focus is on a

[131] Levy, A. and Bechtel, W (2013)
[132] Bich, L. (2012).
[133] In Booderg et al (2007)
[134] A full defense of mechanistic explanation in the sciences is beyond the scope of this paper. For a more detailed discussion, see Bechtel, W., & Richardson, R. C. (1993/2010) and the many sources and discussions referenced there. There is a specific issue of reconciling the mechanistic account of explanation with the information theoretic account of autonomy described in Bertschinger (2008) and discussed briefly in Chapter 2; Ladyman and Ross (2007) go some way towards providing such an account, but this is an avenue warranting further research and discussion.

methodological claim about the role of abstraction in mechanistic explanations. The core of the view is that "mechanistic explanations address organized systems"; the paper argues that an abstracting to the level of "patterns of connectivity" among components is often necessary to explain and understand the organization of a mechanism. The paper proceeds to demonstrate how these abstractions can be usefully represented with the tools of graph theory. By grounding the organization of mechanisms in the abstraction of connectivity, LB hope to motivate a "new mechanistic philosophy" might treat the issue of organization more directly with these theoretical tools in the future. To that end, they offer the following formal characterization of organization:

> "... we will say that given some effect or behavior, a system is organized with respect to that effect or behavior if (a) different components of the system make different contributions to the behavior and (b) the components' differential contributions are integrated, exhibiting specific interdependencies (i.e., each component interacts in particular ways with a particular subset of other components). In our view, both conditions are necessary, although both may be met to varying extents, resulting in different forms and degrees of organization." (243-244)

LB make good on the claim to different *forms* of organization in the examples of feed forward loops in *E. coli.* LB provide three distinct examples of such loops, which Alon (2007a) calls "network motifs". Although each motif is composed of the same quantity of nodes and relations, the particular type of nodes and ordering of relations between them are distinct in each form. These distinctions in organizational structure make a difference for the behavior of the system, and the network motifs successfully capture those dynamics while ignoring unnecessary detail about the operation of components. Because the explanation is abstract, the patterns it describes generalize to any system instantiating the pattern of connectivity, and thus find explanatory purchase in the scientific quest to "unify diverse phenomena".

However, LB give absolutely no clue, at least in this paper, about how one might characterize differences in *degree* of organization, or what role a difference in organizational degree might have in a mechanistic explanation. The three examples from *E. Coli* presumably represent differences in

organizational form between cases of approximately equivalent organizational degree, since each network has equivalent numbers of nodes and relations, albeit arranged in distinct formal structures. To say that two networks differ in their degree of organization presumably implies that one network is *more or less* organized than the other, but we are left without indication of what this difference might be.

LB explicitly treat the idea of degrees of *abstraction*, but there's no reason to think (and LB do not suggest) that the degree of abstraction is related to the degree of organization. Abstraction is a method of "leaving out the details" to focus only on those issues relevant to the phenomenon being explained; an explanatory model is more abstract if it leaves out more details. LB convincingly argue that some abstraction is necessary for capturing the organized features of a system. However, a more abstract model won't necessarily provide a better explanatory model for representing an organized system. In fact, too much abstraction might leave out important details necessary for understanding precisely how a mechanism is organized, leaving an abstraction explanatorily useless. A description of a dog given in terms of its anatomical components reveals a higher degree of organization in the system than a description of the dog as composed of "some stuff". The latter is more abstract, but it is also less helpful in explaining the behavior of the system. Organizational degree, then, isn't simply a matter of how abstractly a system has been described; an organizational description must be given at the *right* degree of abstraction-- the one that is sufficient for providing a mechanistic explanation of the phenomenon in question.

So instead, let's look to LB's formal definition of an organized system for any clues as to how we might cash out the claim of organizational "degrees". Part (a) of their definition involves two distinct kinds of "difference" each of which might potentially make a difference to the degree of organization. First, organizations have different components, and second, the components play different causal (or

functional)[135] roles-- in other words, components come in different types. But on reflection, neither of these differences are sufficient for adequately characterizing organizational degree. Increasing the number of components doesn't necessarily make a system more (or less) organized; a shoebox containing 5 buttons isn't (necessarily) any more organized than the same box with 10 buttons, even though the latter has more components than the former. LB aren't simply interested in the quantity of components, of course; what matters on their account of organization is the differentiation of causal (or functional) role among components. So perhaps more component *types* corresponds to a greater degree of organization: not just the total number of components, but the diversity of causal roles played by components of different types. But that's also clearly false; a shoebox containing of a button, a rat, and a snow globe isn't (ceteris paribus) any more organized than one composed of three type-identical buttons, even though each component in the first box has a distinct causal profile contributing differently to the mechanics of the overall system. Each shoe box differs in organizational form, but these differences seem to make no difference to organizational degree. So we're left to conclude that neither aspect identified in part (a) of LB's definition of organization will be adequate to ground the difference in organizational degree.

Part (b) appears more promising, but runs into deeper worries. LB claim that the components of an organized system are "integrated", which they describe in terms of "exhibiting specific interdependencies (i.e., each component interacts in particular ways with a particular subset of other components)". The claim that the degree of organization corresponds to the degree of integration among components seems to pass initial muster; "more integration = more organization" looks like a plausible enough slogan to endorse. But what feature of a system is being identified when we describe

[135] LB set aside the distinctions between causal and functional interpretations of network motifs. Their emphasis is on the causal roles played by components, but they also recognize (in footnote 7) that these causal roles are expected to perform functions for the system, in the sense of Cummins (1975). I will return to this distinction later in the essay, but for the moment follow LB in leaving both functional and causal interpretations of network motifs open.

the integration among components, and more importantly, when is one system more integrated than another? LB say that integration is a matter of the interdependence of components, described both in terms of the relations (or types of relations) between components, and in terms of the subset of components with which it holds those relations. Both these features of interdependence appear open to differences in degree, with some system potentially having greater or fewer numbers of relations (or relation types), and standing in these relations to larger or smaller subsets of components. Conveniently enough, network theorists call the number of components to which a node is connected the "degree" of that node; the sum of the degrees of each node in a network is called the degree of the network. The degree of the network equals twice the number of edges in the network, since the sum of degrees of all the nodes should count each edge twice[136]. Therefore, a network with more edges has both higher "degree" in the network theoretic sense, and greater interdependence in the terms LB describe. So the proposal appears to be that the degree of a network, understood as twice the number of edges, has some bearing on the degree of organization, understood in terms of the interdependence among components. But on reflection, the degree of a network isn't itself sufficient to account for differences in degree of organization.

To demonstrate this conclusion, we might look at examples of networks with differences in degree (understood in the terms LB provide), and then determine whether these differences make a difference to the degree of organization. This would be the same strategy with which I quickly dismissed the aspects of organization described in part (a) as relevant to the issue of organizational degree. But perhaps these distinctions are too subtle to make confident assessments of their relative degree of organization, especially given that the exact content of this claim has yet to be clearly established. So instead, I'll attempt an example of the opposite sort. In the next section, I'll propose some simple networks with identical degree, but where we nevertheless judge the degree of organization to be

[136] This is known as the handshaking lemma.

distinct. From this it will follow that organizational degree cannot simply be understood in terms of integration, at least as LB have described it. However, my argument is not merely critical. Although these toy examples are deliberately simple, they will help introduce some conceptual tools for analyzing networks that will allow us to focus precisely on the issue of organizational degree and the role it plays in a mechanistic explanation.

4.4 Network degrees in heaven and hell

The example I'm offering is not novel, but is drawn from a parable that appears to be common to many traditions around the world. I've not found any reference to it within the philosophical literature; I originally encountered the parable in a book of logic puzzles as a child, where it was presented as follows:

An old man is curious to know his fate after death, and is miraculously granted vision of the heaven and hell that potentially awaits him (bear with me). He views hell first and sees overflowing tables of delicious looking food, surrounded by the souls of people who appear to be in the grips of starvation-induced agony. On further inspection, he notices that each individual is equipped with oversized pairs of chopsticks that run 6 feet long. These chopsticks are attached to the souls in such a way that any attempt to eat the food will require using the chopsticks. The old man watches one soul after another desperately attempt to coordinate the movement of food into their mouths, and upon inevitably failing, he watches in pity as they cry out in the madness of pure frustration. Having contemplated the torture of hell, the old man then turns to inspect heaven. He is surprised to see a very similar scene: tables of delicious food surrounded by people equipped with 6ft long oversized chopsticks. The old man soon sees soul after soul deftly pick up the food with their chopsticks and happily offer it to their neighbor, whose mouth they could easily reach with their curious extended limbs. Having understood the fates that await him, the old man asks one more time to visit hell and perhaps spare its souls their misery. "Just share with your neighbor, and he will share with you, and

everyone can eat!" the man pleas to the first tortured soul he meets in hell. "Share? With him?" the soul replies. "Why, I'd rather starve."

The parable is meant to make a moral point about sharing, of course, and to reinforce a conception of the afterlife that is both specious and comforting, but neither issue is of interest here. However, I find the formal structure of the networks described in this parable to be useful for getting clear on the issue of integration and distinguishing it from other aspects of an organized network. The parable explains the functional bliss of heaven, in contrast to the dysfunctional torture of hell, in terms of the cooperation among its members: heaven is more organized than hell. This explanation is given at a level of abstraction appropriate for understanding the organizational structure of these populations, in the spirit of LB's discussion of the role of abstraction in explanation. This will allow us to give a precise formal characterization of the differences between heaven and hell, in order to test the proposal under consideration: that organizational degree corresponds to the degree of a network. Although these toy models are obviously far simpler than any of the natural models discussed in serious scientific research, carefully treating differences in the toy case will help us clarify the relationship between integration and degree of organization. Our goal in what follows will be to formally specify the networks of heaven and hell in order to find where they differ in organizational degree. Since heaven and hell differ in a number of ways, our strategy will be to specify these networks so they are as alike in organizational form as possible, while retaining the difference in organizational degree necessary for interpreting the parable. Towards those ends, I offer the following formalization.

Consider first the network Hell, $L= (V_L, E_L)$, where V_L is a set of nodes representing each soul in hell, and E_L represents the set of feeding relations between them. We can define E_L formally by a function $f_L : V_L \rightarrow V_L$ such that $u \in V_L$, $f_L(u) = u$. This function expresses the identity relationship and corresponds to the fact that in hell people only attempt to feed themselves. In a network diagram, this might be represented with a self-directed loop at each node. However, the identity feeding relation is

unsuccessful. So to simplify our depiction of hell, we will exclude these unsuccessful self-directed loops. Hell, in this representation, is a completely disconnected network (the edgeless graph) that might be depicted as follows for the case where $|V_L| = 9$:

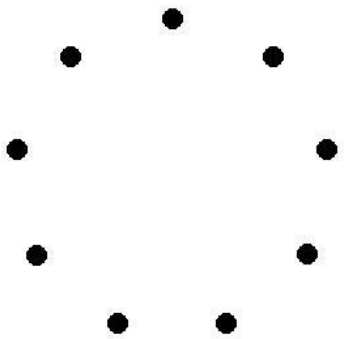


Figure 1. Hell, L

Next, consider the network Heaven, $N = (V_N, E_N)$, where V_N is a set of nodes representing each soul in heaven, and E_N is the set of edges between them. The parable presents heaven as the situation where everyone feeds each other, though like hell no one can feed themselves. Formally, $E_N = \{ \{u, v\} : u, v \in V_N, u \neq v\}$, which is to say that E_N contains every pairwise set in V_N without fixed points. In a network diagram, this can be represented by a complete graph[137] as in Figure 2 for the case $|V_L| = 9$. A complete graph is fully connected (each node has an edge to every other node) and therefore represents the network of highest possible degree, the exact complement of L. As complements, N and L represent the maximum and minimum degree for a network of equivalent size. Although Figure 2 depicts (n-1) = 8 undirected edges at each node, each of these edges in fact represents two distinct (but symmetrical) directed feeding relations: where $u, v \in V_N$, $f(u) = v$ and $f(v)=u$. So when counting the degree of the network, we should make sure to count each edge in this diagram twice.

[137] For the sake of consistency, our diagrams will continue to assume the number of nodes $n = |VL| = |VN| = 9$. As I argued in part 1, the number of components of a network is not itself a measure of organizational degree. *Ceteris paribus*, assuming $|V_L| = |V_N|$ should not have an impact on the relative difference in organizational degree between networks L and N. So I treat this assumption as benign for our purposes, despite empirical evidence suggesting a larger population in hell.

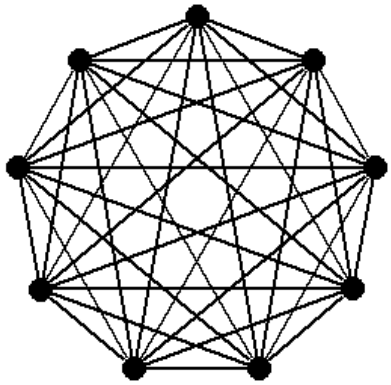
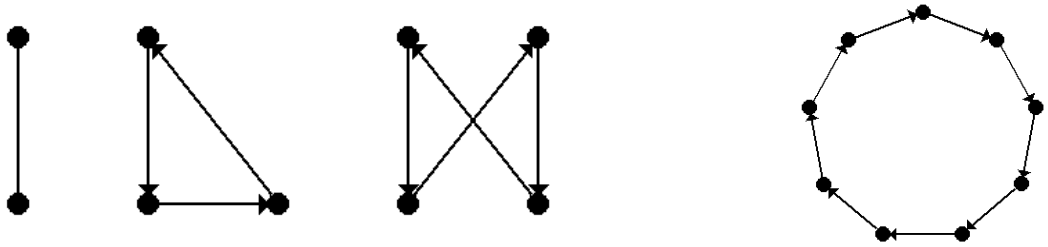


Figure 2. Complete heaven, N.

The degree of N is obviously greater than the degree of L. Specifically, each node in complete heaven $i \in V_N$ has degree $d_i = 2(n-1)$ (= 16 for n=9). Summing over all nodes, this puts the total degree of the network for $d_N = 16*9 = 144$. In contrast, each node in hell has $j \in V_L$, $d_j=0$, since we're excluding self-loops, which puts the degree for the total network $d_L = 0$. From this alone we might conclude that heaven is more organized than hell. Indeed, the second criterion in part (b) of LB's definition identifies the degree of a node as an important element in evaluating its interdependence, and this result seems to support the conclusion.

But consider instead a more restricted version of heaven $R = (V_R, E_R)$, where each person feeds exactly one other person. As long as the entire network is fed and no one feeds themselves, this restriction would keep with the spirit of the parable in distinguishing the more organized heaven from the less organized hell. The restricted set E_R is defined by a function $f_R: V_R \rightarrow V_R$ that is both bijective (one-to-one and onto) and has no fixed points ($u \in V_R$, $f_R(u) \neq u$). A number of configurations satisfy this function, two of which are represented in Figure 3.



a b

Figure 3. Two possible configurations in restricted Heaven, R

Other configurations are possible. In combinatorics, the bijective self-mappings with no fixed points are called the "derangements" of the set, and the number of derangements is represented recursively by !n = (n-1) (!(n-1) + !(n-2)). For the case of n=9, there are 133,496 possible derangements of complete heaven that would satisfy f_R. So there's lots of ways to arrange heaven under this restriction that still has everyone fed and happy.

Whichever of these configuration is picked, however, if it satisfies the edge function f_R, then each node will have exactly two directed edges. This is easy to confirm in both configurations represented in figure 3. Each node is both feeding (outdegree=1) and being fed by (indegree=1) exactly one other node[138], and the total degree for each node is the sum of its indegree and outdegree. Therefore, every configuration of R will have degree $d_R = 2n$, or 18 for n=9. Earlier we claimed that R is more organized than L, and indeed $d_R > d_L$, supporting the proposal under consideration. But we've shrunk the difference in degree to a mere 12.5% of the difference between N and L, so we're making progress! We might be tempted to stop here and ask whether we would accept the claim "R is 12.5% as organized as N", but this would just distract from our goal, which is to generate an example of equivalent network degree which nevertheless differs in organizational degree. One last adjustment to our parable should do.

Consider, finally, a more relaxed Hell, X = (V_X, E_X). Relaxed hell does away with the silly chopstick

[138] In the case where the feeding relationship is symmetric, as in the leftmost component in the network labeled a, the edge is represented as an undirected graph representing two distinct symmetric relations, as in figure 2.

rule, and gives the souls in hell complete and unrestricted access to the bountiful food around them. Nevertheless, in relaxed hell the souls retain their selfishness and jealousy towards their neighbors, and refuse to feed anyone but themselves; presumably the souls in relaxed hell continue to be tortured in other ways so as to maintain their bitter disposition despite their full bellies. E_X will be formally defined by the identity function similar to the definition of E_L, but with the important exception that in X these relations are *successful*, and therefore deserve to be counted among our edges and explicitly represented by self-directed loops in our diagrams, as in figure 4 for the case of $|V_X| = 9$.

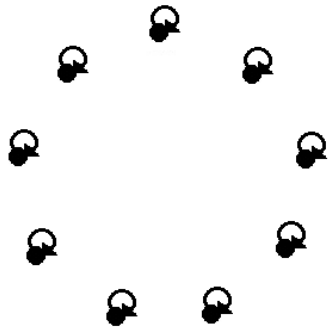


Figure 4. Relaxed hell, X

Since the feeding relation is successful in X, the degree of X will be greater than $d_L = 0$. The degree of the network is the sum of the degrees of all the nodes, which is twice the number of edges. And since $|E_X| = n$, then $d_X = 2n$ (= 18 for n = 9). Moreover, since $|E_X| = |E_R|$, we can conclude that $d_X = d_R$.

In other words, relaxed hell X shares with restricted heaven R exactly those formal features identified by LB as relevant to organization. By assumption, X and R have equivalent numbers of nodes, and all of the nodes in both networks are of the same types. So X and R are alike by the criteria in part (a). Furthermore, X and R also share an equivalent number of (successful) feeding edges between its nodes, and the degree of each node in both networks is 2, giving a total network degree of $d_X = d_R = 2n$; if this is not clear from Figure 4, be sure to count both the indegree and outdegree of each node. In graph theoretic terms, both are directed 2-regular graphs, with each node in both networks having

identical (in- and out-) degrees. Equivalent degrees imply that each node in both networks is connected to subsets of components of equivalent size, thus meeting the criteria in part (b). So X and R share all the relevant formal structures, and in equivalent magnitudes, according to the definition of organized systems provided by LB. If the degree of a network corresponds to its degree of organization, then X and R are equally organized.

Or perhaps not! Closer inspection of the criteria in part (b) of LB's definition reveals that interdependencies are described in terms of interactions between "subset of *other* components". The components in R have relations to each other, and the components of X have relations only to themselves. If "integration" requires that components are connected to components other than themselves, then these networks are not alike in terms of integration of components. Consequently, network degree doesn't capture the sense of "connectivity" relevant to integration. The proposal under consideration appears to have been undermined by the example. One might try to salvage the proposal by insisting that we ignore any self-loops, whether successful or not, when assessing network degree. This interpretation would treat X as an effectively edgeless graph identical in degree to R; the implication is that the nature of the identity relation is irrelevant to the organization of a graph, and that we might effectively disregard the contributions to a network from disconnected or isolated nodes.

But this strong interpretation of integration seems too strong, since we may find cases where isolated nodes nevertheless contribute to the dynamics under consideration. Consider, for instance, the network described in the Bearman et al. (2004)[139] study of the pseudonymous Jefferson High School.

Bearman reconstructed the network of romantic relationships among the high school students by correlating the student's self-reports self reports of their romantic encounters. Quoting from the study: "Figure 2 maps the actual network structure that links the 573 students involved in a romantic or sexual relationship with another student at Jefferson High. Circles denote individual students; romantic

[139] Bearman et al. (2004).

or sexual relations are denoted by connections between nodes. Time is suppressed in this representation." Each student in the survey was asked to list the three most recent romantic relationships, and the representation demonstrates all every link that exists over that period. As this real-world example makes clear, the disconnected components of a network might nevertheless be relevant for assessing the organization of the network. Cases of discontinuity in this network are will be relevant, for instance, in tracking the flow of sexually transmitted diseases among this population.

If organizational degree corresponds to degree of interconnections across distinct nodes in the network, we'd be tempted to assert that the upper left component of the Jefferson High network is the most organized of the bunch. But this doesn't follow. The density of interconnections between some component or other of a network is described as its "clustering". A network like complete heaven N, where node is connected to every other node, is called a "clique", and a network is more or less clustered depending on how close the nodes in a network come to being part of a clique[140]. The upper left component in the Jefferson High network is certainly more cliquish than the other components, but they aren't necessarily more organized.

One final consideration remains, and although it isn't mentioned explicitly in LB's definition it nevertheless is a distinct part of the motivation for their work. What matters for organizational degree is not simply how interconnected the nodes are, but more importantly how well those interconnections contribute to an explanation of the mechanism described by the network. This is to say that the organization of the network only makes sense within the context of a description of the function or phenomenon the mechanism supposedly satisfies. If we (naively) take the "goal" of high school relationships to be about the establishing of coupling relationships among the students, then the isolated components with a few nodes might do just as well at the task as the cliquish community in the top right, so the differences in the densities of interconnections don't make a difference to its

[140] For a fantastic survey of the various modeling techniques for describing clustering patterns in networks, see Fortunato, S. (2010) and the references therein.

organizational degree.

Returning to the networks of heaven and hell, an unstated assumption of the parable is that the populations in both networks have the goal of being fed. For this reason we raised suspicions that relaxed hell, where people can feed themselves and everyone gets fed, violates the point of the parable, because the population's goal are satisfied. However, notice how the explanation for this satisfaction differs between relaxed hell X and restricted heaven R. In X, the satisfaction of the goal of feeding everyone can be explained purely by way of local facts about what each agent does considered in isolation, but in R requires appreciating global facts about the relations among the agents as a whole (that E_R is a bijective mapping). Although alike in network degree, R and X differ in their interconnectedness, which isn't simply a matter of the numbers of connections but a matter of how well those connections contribute to the function of the network considered as a whole mechanism. A network with fewer connections may nevertheless be more organized if those connections are better suited to satisfy the mechanism under consideration.

This last point, about mechanism-relative interconnectedness, brings us back to the discussion of technological networks that has motivated this dissertation. Throughout this essay I've described functions, agents, tools, and their relations in task-relative terms. This perspectivalism is implicit in LB's discussion of abstraction, and is central to any account of organization that treats the phenomenon as an emergent property of coordinated, interacting components. The structures relevant for an explanation of the organization of some system depend on the sort of explanation we seek, and what purposes we want to put these explanations to. The claim that machine participants are explanatorily relevant for understanding our technological world is a claim that at the scales necessary for describing the dynamics of our social communities, at least some machines must be understood as participating agents in those communities. There is obviously much more to say about the structure of mechanisms in

natural organizations[141], but I leave this discussion to conclude with some remarks about organized sociotechnical systems.

4.5 The infiltration of the machines.

In the previous two chapters, I defend a conception of machine participation largely through negative arguments against mainstream views of artifacts and autonomous agents. In this chapter, I hope to have supplemented those critical arguments with by introducing network-theoretic concepts and some suggestive examples to provide a constructive glimpse at an alternative picture of technological development. I want to close this discussion with two final examples that, I hope, will provide conclusive support for the significance of the thesis, and why a scientifically integrated philosophical analysis of the sort provided in this dissertation will be useful for the world we are so busy creating.

The first comes from Ciulla et al (2012),[142] and involves a study of Twitter activity used to predict the outcome of the season finale of an American Idol episode. Using a very simple predictive model and data publicly available from the Twitter firehose, the study made a prediction on the winner of the audience-voted results, and indeed their model was confirmed when the results were announced. Their study ends with a word of warning about the use of readily available public data as a predictive instrument in social events, especially as it pertains to events with large financial interests. They continue: "On a more general basis, our results highlight that the aggregate preferences and behaviors of large numbers of people can nowadays be observed in real time, or even forecasted, through open source data freely available in the web."[143] For our purposes, the important result from this study is the incredible success and accuracy of using the sociotechnical networks we've recently constructed (like Twitter) to obtain a picture of human behavior at a level of abstract analysis that allows for predicting

[141] See Ladyman and Ross (2007), Ladyman and van Frassen (2007) and the debates that followed.
[142] See more discussion of this study in Chapter 4.
[143] Ciulla et al (2012)

and explaining large scale social phenomena in ways that have only recently become possible. Ciulla's result is not particularly unique or surprising[144], as similar techniques using similar social networking services have recently created an explosive interest in data modeling and statistical analysis techniques that have come to be described as the "digital humanities".[145] The predictive and explanatory power of these analyses on our sociotechnical systems provide compelling reasons for adopting the network-theoretic methods that have become widely accepted in the field, and give support to the views defended in this chapter.

But these results take on a more intriguing character when compared to another area of research tracking the influence of social bots on Twitter[146]. Boshmaf et al (2011) find that social bots can effectively infiltrate social networks and mimic the activity of real users without detection; surprisingly, it found that 80% of the bots used in the study had evaded detection over the course of the study. Mitter (2014) demonstrates the real influence these bots can have on the dynamics of social networks. Both studies conclude with the expected remarks about privacy and security in the face of growing, coordinated infiltration of social networks by bots.

Placed alongside with the Cuilla study and the general trend of using social networks to predict and explain large scale social phenomena, an obvious conclusion presents itself: social machines have infiltrated the very networks by which we attempt to model, explain, and predict aggregate human behavior. Automated programs have set up shop in those communities, behaving for all the world as a participant coordinating their activities with the rest of the community, drifting along with the social tide and passing as one of us without detection. This is precisely the world Turing predicted we would have achieved by the century's end. Despite the many reasons for think such a world would never occur,

---

[144] For a discussion of the techniques and challenges in this body of research, see Ratkiewicz, J.(2011), Lehmann, J. (2012), Castillo (2011), and the references there.

[145] A term I take to include work in digital ethnography, digital sociology, etc. See Schreibman et al. (2008)

[145]

[146] See Boshmaf, Y. (2011), Hwang (2012), and Mitter (2014)

we’ve gone and built it anyway. Mostly without noticing that we had done it at all.

Chapter 5

Gaming the Attention Economy[147]

5.1 Abstract

The future of human computation (HC) benefits from examining tasks that agents already perform and designing environments to give those tasks computational significance. We call this *natural human computation* (NHC). We consider the possible future of NHC through the lens of Swarm!, an application under development for Google Glass. Swarm! motivates users to compute the solutions to a class of economic optimization problems by engaging the attention dynamics of crowds. We argue that anticipating and managing economies of attention provides one of the most tantalizing future applications for NHC.

5.2 Natural Human Computation

Human computation (HC) involves the creation of mixed organic-digital systems to solve difficult problems by outsourcing certain computational tasks to the human brain. However, we can distinguish between HC approaches that require a user to engage with a specific (and arbitrary) program or system, and HC approaches that simply leverage a user's normal activity to compute the solutions to complex problems. We call this latter approach *natural human computation* (NHC). An instance of HC is *natural* when the behavior necessary for carrying out the proposed computation is already manifest in the system.

Eusocial insect colonies are models of natural computation (see Gordon, 2010; Michelucci, 2013). The information processing potential of ant colonies emerges from the small-scale, everyday interactions of individual ants: everything individual ants do is computationally significant, both for the

[147] Cowritten with Jonathan Lawhead. Reprinted with permission from Michelucci, P. (2013). *Handbook of human computation*. Springer.

management of their own lives and for the colony's success. This alignment between individual and colony-level goals means that ant colonies need not direct the behavior of individual ants through any sort of top-down social engineering. The queen issues no royal decrees; insofar as she has any special control over the success of the colony, that control is a product of her influence on individual colony members with whom she comes into contact. The sophisticated information processing capabilities of the colony as a whole are a product of each ant obeying relatively simple local interaction rules--those local interaction rules, however, allow an aggregate of ants to influence each others' behavior in such a way that together, they are capable of far more complicated computing tasks than individual colony members would be on their own. Crucially, the computational power of the colony *just is* the concerted action of individual ants responding to the behavior of other ants: any change in the colony's behavior will both be a result of and have an impact on the behavior of colony members. In this sense, natural ant behavior is both *stable* and *natural:* the computing activity of the colony can't *disrupt* the behavior of colony members out of their standard behavior routines, since those standard behavior routines *just are* the computing activity of the colony. The stability of this behavior can in turn support a number of additional ecological functions. The regular harvesting of individual bees not only supports the activity of the hive, but also solves the pollination problem for flowers in what we might call "natural bee computing"[148] which piggybacks on the behavior. NHC approaches take these natural models of computation as the paradigm case, and seek to implement similar patterns in human communities.

We have sketched a definition for NHC in terms of *stable* and *disruptive* computation, and turn now to discuss these concepts directly. Disruptive computation requires a *change* in an agent's behavior in order to make their performance computationally significant. Human computation is increasingly *stable* as its impact on agent behavior is reduced. Describing an instance of human computation as "natural" is not itself a claim that the *human activity* is stable or disruptive, since NHC

[148] Of course, bees and flowers achieved this stable dynamic through millions of years of mutualistic interaction; as we discuss in section 4, we expect any HC technique to require some period of adaptation and development.

techniques can be used to extract computationally significant data in both stable and disruptive contexts. Rather, describing an instance of HC as natural makes the more limited claim that the computation in question was not *itself* a source of disruption. We introduce the vocabulary of stability and disruption to clearly articulate this aspect of NHCs.

It may be instructive to compare NHC and gamification (Deterding et al., 2011; McGonigal, 2011) as strategies for human computing. Gamification makes an HC task more palatable to users, but often alters user behavior in order to engage with the computational system. In contrast, NHC systems transparently leverage existing behaviors for computation. For instance, reCAPTCHA (von Ahn et al., 2008) repurposes an existing task (solving text-recognition puzzles to gain access to a website) to solve a new problem (digitizing books for online use). This pushes HC to the background; rather than explicitly asking users to participate in the solution of word recognition problems, it piggybacks on existing behavior. Gamification is not always disruptive in the sense used here; in some cases described below gamification techniques can serve to *stabilize* (rather than *disrupt*) the dynamics of systems to which they are applied. This suggests that we need a more robust vocabulary to map the conceptual territory.

Michelucci (2013) distinguishes between "emergent human computation" and "engineered human computation." Emergent HC systems analyze uncoordinated behavior from populations to do interesting computational work, while engineered HC systems might be highly designed and coordinated for specific computing needs. We see natural human computation as a concept that is complementary to but distinct from Michelucci's distinction. The defining characteristic of NHC is the potential for extracting additional computational work from human activity without creating additional disturbances in that behavior. This definition makes no assumptions about the degree to which these behaviors have been designed or coordinated for particular computing functions. In fact, we assume that natural human behavior involves organizational dynamics that cut across Michelucci's distinction. NHC systems like Swarm!, described in Section 3 below, can be understood as a method for discerning natural

organizational patterns as a potentially fruitful source of human computation.

We're thinking about NHC in terms of the impact a computing task has on the behavior of its computers; NHC tasks introduce minimal disruptions to existing behavior. In contrast, Michelucci's distinction isn't concerned with the impact HC has on its agents. Rather, it is concerned with the performance of the computing task in question. Emergent cases of computing are where the goal is best accomplished by passively analyzing agents for specific computational results, more or less independent of other aspects of their behavior. Engineered systems require increasingly coordinated activity to achieve computational results. For these reasons, we consider Michelucci's distinction to be a system-level or "top-down" perspective on computing tasks, while the stable/disruptive distinction is an agent-level or "bottom-up" perspective on the same tasks. Or to cast the issue in techno-phenomenological terms: Michelucci is taking a designer's perspective on human computing, where purposes (functions, tasks, goals, ends) are *imposed* on a computing population; on the other hand, we're interested in the user's perspective, where the generation and pursuit of purposes is a constitutive aspect of one's ongoing committed engagement with the world.

It is worth reiterating that the sense of "natural" being articulated cuts across the categories represented in Table 1 below. We can think of these categories as defining the axes of a continuous space of possible computing systems. Claiming that a given system is emergent and disruptive (for instance) is to locate within this space. However, claiming that a given instance of human computation is *natural* is to point out a very different sort of fact about the system. In the context of human computation, *naturalness* is something like an indexical, describing words with use-relative content like "here" or "now." Rather than giving an absolute location in the space defined by the distinctions discussed above, calling an instance of HC "natural" is to assert a fact about the HC system *relative* to the current state of the computational substrate. A NHC might be engineered, emergent, disruptive, or stable to some greater or lesser degree; the ascription of naturalness depends only on a comparison

between the system's state *now* and the state that would be necessary for performing the desired computation. These distinctions can be more clearly illustrated if we consider a few representative examples. Consider the following cases:

	Stable	**Disruptive**
Emergent	American Idol predictions	Yelp
Engineered	Zombies Run	FoldIT

Table 1.

Emergent/Stable: HC systems are emergent when they exploit uncoordinated behavior in a population, and they are stable when that computing goal is met without further disruption. reCaptcha has already been mentioned as an example of HC that falls in this quadrant. A more illustrative example can be found in Ciulla et al. (2012), which describes modeling approaches to the Twitter datastream that successfully anticipate the results of a recent American Idol voting contest. In this study, users Tweeted their thoughts on the contest of their own accord[149], without coordination and independently of their potential use in predictive speculation, and so meets the definition of emergent. Solving the prediction task required no additional input from the users beyond this existing social behavior, and so also meets the definition of stable.

Engineered/Stable: Engineered computing tasks are highly coordinated and designed for specific computing purposes. These designs can be stable in our sense when the computation fits existing patterns of behavior rather than creating new ones. BOINC's successful @HOME distributed computing projects (Anderson 2004) are familiar examples of stable computing strategies, using spare processor cycles for useful computational work without occupying an additional computational

[149] We ignore for the sake of the example any potential feedback from advertising or other systems that reinforce tweeting behavior surrounding the American Idol event.

footprint. For a more explicitly gamified example, consider the 2012 exercise motivation app called "Zombies Run"[150]. Zombies Run (ZR) is designed to work in tandem with a player's existing exercise routine, casting her as a "runner" employed by a post-apocalyptic settlement surrounded by the undead. The game's story is revealed through audio tracks rewarding player for gathering supplies, distracting zombies, and maneuvering through the dangerous post-apocalyptic wasteland, all accomplished by monitoring a few simple features of the user's run. The app motivates runners to continue a routine they've already developed, using tools already appropriated in that behavior; the app isn't designed to help users to start running, it is designed to help them *keep* running. This is a defining feature of engineered/stable systems: while they are the product of deliberate design, the design's primary effect is to reinforce (rather than alter) existing patterns of behavior. While ZR players aren't (necessarily) performing any particularly interesting computing function, the app provides a clear example of how a highly designed, immersive app can nevertheless be stably introduced into a user's activity.

Emergent/Disruptive: A computational state is *disruptive* when implementation would involve a significant reorientation of the behavior and/or goals of the agents under consideration. This can occur in emergent computing contexts where individuals are acting independently and arbitrarily. Yelp.com is a popular web-based service that compiles crowd-sourced reviews of local businesses and services. These reviews are used to compute a rating of a given service based on search criteria. And indeed, solving this computing problem itself changes the activity of the population: Luca (2011) finds that the one-star rating increase amounts to a 5-9 percent increase in revenue. In other words, the self-directed, emergent activity of Yelp reviewers is disruptive to the behavior of the dining community, effectively redirecting a portion of them to services with higher ratings. It may be supposed that Yelp's disruptive status is a consequence of feedback from the HC system being used to guide the decisions of future

[150] From the UK-based Six to Start. https://www.zombiesrungame.com/

diners. However, Zombies Run provides an example where feedback on HC behaviors can reinforce those behaviors with little disruption. This suggests that Yelp's economic impact involves more than providing feedback on the HC task; it reflects something about the specific computations performed by the system. We will return to this point in section three.

Engineered/Disruptive: FoldIT is a puzzle-solving game in which the puzzles solved by players are isomorphic to protein folding problems (Khatib et al. 2011). FoldIT is a paradigm case of gamification: it makes a HC task more palatable to the users, but significantly disrupts their behavior in the process by demanding their focus on the game. FoldIT is engineered in the sense that the task has been deliberately designed to provide computationally significant results, and disruptive in the sense that the task is a departure from the behavior in which players otherwise engage.

The above examples are offered in the hopes of making clear a complex conceptual landscape that serves as the backdrop for the discussion of natural human computing. A full discussion of the dynamics of purposive human behavior is beyond the scope of this paper, but we understand our contributions here as a step in that direction. Notice that the above examples contain no direct appeal to "naturalness", since the degree of naturalness for some HC process may vary between individual users with distinct behavioral profiles. Using Yelp in deciding on some service, or using ZR to motivate your run, will integrate naturally into the usage patterns of some users and may be more disruptive in the lives of others. Or in the terms introduced earlier, "naturalness" can be understood as an index of a user's state. Despite the perspectival dimensions of the term, we can talk sensibly about designing natural human computing systems that leverage existing HC work in minimally disruptive ways. We turn now to describe a NHC system that demonstrates these features.

5.3.1 Introducing Swarm!

Swarm!, an application under development for Google Glass[151], is an implementation of NHC methods for solving a class of economic optimization problems. Swarm! uses the GPS coordinates of players to construct a location-based real time strategy game that users can "play" simply by going about their everyday routines. Individual cognitive systems have limited resources for processing data and must allocate their attention (including their movement through space and time) judiciously under these constraints. Therefore, we can interpret the data gathered by Swarm! as a NHC solution to the task of attention management: Swarm! generates a visualization of aggregate human activity as players negotiate their environments and engage objects in their world.

The Swarm! engine is designed as a crude NHC application: it's a game that's played just by going about your normal routine, frictionlessly integrating game mechanics into a player's everyday life. Swarm![152] simulates membership in a functioning ant colony, with players assuming the role of distinct castes within one colony or another. Players are responsible for managing their own resources and contributing to the resource management of the colony. Swarm! data is visualized as colorful trails on a map card presented on request to the Glass display in order to engage the resulting behavior. These trails are designed so they cannot be used to locate or track any individual uniquely. Instead, we're interested in the broader patterns of behavior: where do players spend their time? When is a certain park most likely to be visited? When and where do players from two different neighborhoods cross paths most frequently?

5.3.2 Swarm! mechanics

Ant behavior is coordinated through purely local interactions between individuals and a shared

[151] Glass is a wearable computer designed and manufactured by Google. The Glass headset features a camera, microphone with voice commands, optical display, and a touch-sensitive interface. It duplicates some limited functions of a modern smartphone, but with a hands-free design. Fig. 1 depicts a user wearing a Google Glass unit.
[152] Complete game bible can be found at http://www.CorporationChaos.com

environment without any central direction (Dorigo, 2000). Individual ants exchange information primarily through direct physical contact and the creation of pheromone trails. Pheromone trails, which can be used to indicate the location of resources, warn of danger, or request help with a tricky job, are temporary (but persistent) environmental modifications laid down by individuals that help ants coordinate with each other and organize over time to manage colony needs.

Swarm! adopts the pheromone trail as its central mechanic. By moving around in physical space, players lay down "trails" that are visible through the in-game interface as colorful lines on a map. These trails encode context-specific information about the history and status of user interactions around a location. Just like real-world ants, Swarm! trails are reinforced by repeated interaction with a region of space, so the saturation of trails in a particular location represents the degree of activity in that location. Trails also encode some information about in-game identity, but the focus of Swarm! is on impersonal aggregate data and not unique player identification. Since trails are semi-persistent and fade slowly with time, the specific time that a player passed a location cannot be deduced by looking at the map. Players also have the option to define "privacy zones" around their homes and other sensitive areas where Swarm! data collection is prohibited.

Swarm! gameplay is styled after many popular resource collection games, with central goals revolving around finding enough food to stay alive, disposing of trash ("midden"), and defending the colony from incursions by rivals. However, Swarm!'s central innovation is its emphasis on self-organized dynamic game maps and frictionless player interaction. Player interactions result primarily from trail crossings: when one player crosses the trail laid down by another player, an appropriate context-dependent event is triggered. Note that this activity does not require players to be present simultaneously at one location. Trails laid down by users decay gradually over time, and require reinforcement to sustain. Thus, crossing the trail of a rival ant means that ant (or possibly several ants from the same colony) has reinforced this trail within the decay period. In other words, all player activity

is rendered on the map as “active” and will trigger engagements and events specific to those interactions.

Players also have caste-specific abilities to augment the structure of the game map. These abilities are triggered by more in-depth interaction with a location--for instance, spending an extended amount of time in the same public place, or taking some number of pictures of an important game location. Each caste has a unique set of strengths, weaknesses, and abilities that affect the range of in-game options available to the player. These augmentations can provide powerful bonuses to members of a player’s colony, hinder the activities of rivals, or alter the availability of resources in the area. Strategic deployment of these abilities is one of the most tactically deep and immersive aspects of Swarm! gameplay.

For illustration, consider the following in-game scenario (Fig 5). Suppose a player (call her Eve) consistently moves through territory that is controlled by an enemy colony--that is, she crosses a region that is densely saturated with the trails of hostile players. Moving through this region has a significant negative impact on Eve’s resource collection rate, and unbeknownst to Eve (who doesn’t like to be bothered by game updates) this penalty has been adversely affecting her contributions to her colony for weeks, keeping her at a relatively low level than where she might be otherwise. However, suppose that one day Eve decides to actively play Swarm!. Upon downloading the latest game map she observes the impact this region has had on her collection rate. Swarm!’s game mechanics reward this attention to detail, and allow Eve to do something about it. When Eve photographs the locations that are controlled by a rival colony, she creates an in-game tag that calls attention to her predicament and provides caste-specific in-game effects that potentially offset the impact of the rival colony’s trail. In other words, her action (taking a picture) has produced an in-game structure that warps the map and partially ameliorates the penalty that she would otherwise suffer. This in-game structure might attract other active players to the territory to build more structures that further magnify these changes. In this way,

close attention to (and interaction with) the game map is rewarded, while casual players are still able to contribute meaningfully to the overall game dynamic.

This reveals an important aspect of Swarm! related to the distinctions drawn in Section 1. Although the game is designed to passively harvest aggregate user behavior, it also incentivizes the curation of that data allowing for active user engagement. Thus, some users may experience Swarm! as unobtrusive and stable, with computation occurring largely in the background, while others may enjoy significant disruptions as they actively play the game. Moreover, the two might interact with each other through in-game mechanics around shared spaces without either player being aware of the other's presence. When Eve tags a highly trafficked area of the map with her picture, she is highlighting an attractor[153] in *both* the physical space and the game space. Those attractors emerge naturally in the behavior of some Swarm! players, and Eve's active engagement with the trails further augments the map to highlight the relevance of those attractors. These attractors can in turn coordinate others to further document and engage an area, filling out the digital profile of regions that are of central use in human social behaviors, and effectively turning Swarm! players into an engineered team of self-directed, self-organized content curators. Every Swarm! player's behavior is thus influenced both by the structure of the game map, and the structure of the game map is influenced by the behavior of Swarm! players. However, since the initial structure of the Swarm! game map is dictated by the antecedent behavior of Swarm! players, this mechanic only serves to reinforce extant patterns of behavior.

[153] An *attractor* is just a location or state in a system toward which nearby states or locations tend to be "sucked." Minimum-energy states in mechanical system are commonly attractors. For instance, in a system consisting of a marble confined to the inside of a mixing bowl, the state in which the marble is at rest at the bottom of the bowl is an attractor: no matter where you start the marble, it will eventually end up at rest at the bottom of the bowl. For an accessible introduction to the language of attractors and dynamical systems theory, see Strogatz (2001) and Morrison (2008).

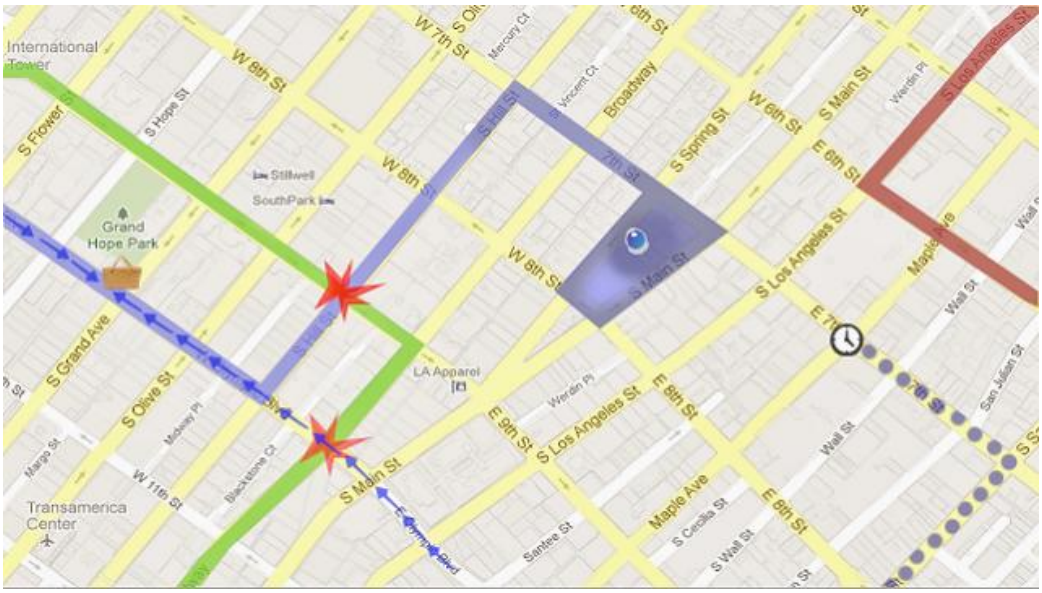


Figure 5. Our player Eve (indicated with the blue dot at Spring and 7th streets) considers a regular interaction with a hostile colony at the corner of Hill and 9th streets.

The resulting model highlights patterns of natural human behavior that can be directly harvested for computational work. For instance, consider the problem of locating a working electrical outlet at the airport[154]. Traditional resource distribution structures (like the financial markets or public regulatory structures) have until now failed to provide enough incentive to curate a digital outlet location map for wide public use, despite its potential value to both customers (who may wish to charge their electronics while they wait for a connecting flight), and the airport businesses (who might be able to draw customers and control the flow of airport patrons by advertising their location). Online databases like Yelp work well for services that have existing advocates, like restaurant owners, who can represent those interests by responding and reacting to Yelp reviews, but little incentive exists for a curation task like this. On the other hand, with suitable resolution Swarm! provides an immediate visual representation of the activity of airport patrons that allows for intuitive predictions about where the outlets might be: look for clustering behavior near walls. Moreover, Swarm! rewards active players for tagging public spaces with pictures and notes that fill in details of the interaction history at that location.

[154] Credit goes to Robert Scoble for raising the example during a recent conversation about Swarm!.

The result is an NHC method for computing a solution to the problem of finding electrical outlets without the need for natural advocates or market representation to explicitly engineer this behavior.

This example has Swarm! players uncovering the use-value of objects which have been under-represented by other records of social value, and it has accomplished this without creating any additional demand on social behaviors. Perhaps a close analog is the use of GPS requests for identifying traffic congestion (Taylor, 2000), but the game mechanics of Swarm! generalizes the approach for a broad range of human activities. We turn now to a general discussion of the strategies described above.

5.3.3 NHC Applications of Swarm!?

Consider the mechanic described in Section **5.3.2** for modifying the game map by taking and tagging pictures. A strategically-minded Swarm! player will not use this ability at just any location (Rashid, 2006; Ames & Naaman, 2007); rather, she will study the structure of local trails over the course of a few days, and engage with the map in a tactically-optimal location--that is, a location that already experiences heavy traffic of the right sort. In this way, the Swarm! map will become a fairly detailed representation of patterns of player engagement with the real world; locations that are naturally highly trafficked will become increasingly important, and thus increasingly saturated with trails and in-game structures.

The fact that interesting locations in the game tend to mirror the interesting locations in the real world is central to Swarm!'s design. While Swarm!'s mechanics might well have some influence on the behavior of more strategically-minded players, that influence will *remain* a mirror of the aggregate pre-game behavior of the community, and thus a useful starting point for NHC data collection about use behavior. Ingress, a somewhat similar augmented reality game developed by Niantic Labs for Android mobile devices (Hodson, 2012), makes for an instructive contrast case. Ingress features two in-game "teams" (Enlightened and Resistance) involved in attempts to capture and maintain control of "portals," which have been seeded by Google at various real-world locations. Players take control of a portal by

visiting the location (sometimes in cooperation with other players), and remaining there for a set amount of time. Players may also "attack" portals controlled by the opposing team through a similar location-based mechanic.

Notice the difference between tracking the behavior of Ingress players and tracking the behavior of Swarm! players. Despite both games featuring similar location-based mechanics, the fact that Ingress' portals--the significant in-game attention attractors--have been seeded by the game's designers renders the activity of Ingress players a poor proxy for their natural, out of game behavior, and thus a poor proxy for NHC data collection. In contrast, Swarm! players create the structure of the map themselves, and the strategically optimal approach to modifying it involves reinforcing existing patterns of behavior. The structure of the Swarm! map reveals at a glance sophisticated facts about the natural attention patterns of Swarm! players. It is this fact that makes Swarm! an important first step toward a mature NHC application.

Transitioning Swarm! from a NHC-oriented game to a real NHC application will involve tightly integrating Swarm!'s mechanics with real-world tasks. We suggest that Swarm!'s existing mechanics might be easily tied in to a service like Craigslist.org. Craigslist is a popular and free web-based service facilitating the exchange of goods and services that run the gamut from used cars and furniture to prospective romantic encounters--all of which are organized geographically and easily searchable. The Swarm! platform, with its built-in mechanics for tracking location, activity, and experience could serve as a platform for visualizing Craigslist service requests and evaluating the results of the transaction. If successful, such a system would allow for a self-organized, entirely horizontal resource and labor management system for its users. Such integration would be a large step toward turning Swarm! into the sort of robust economic HC application that we discuss in Section 5.4.

Consider the following hypothetical future in-game scenario: Eve, our intrepid player from Section 5.3.3, has access to a Craigslist-like service integrated with an advanced version of Swarm!, and

this service informs her (on request) about posts made by other players in her immediate geographical region. With access to this information, Eve can decide whether or not to accommodate the requests of other players in her vicinity. Suppose, for instance, that Eve notices a posting near her home base requesting a 40 watt CFL light-bulb to replace a bulb that just burned out. Eve was targeted with the request because her patterns of behavior repeatedly cross paths with the requesting user; depending on how sophisticated the service has become, it might even recognize her surplus of light bulbs. In any case, Eve knows that she has several matching bulbs under her kitchen sink, and considers using the bulb to gain experience and influence within Swarm!. Eve notices that the specified drop point is on her way to work, and agrees to drop the bulb by as she walks to the subway. Perhaps the drop-off is coordinated by each party taking a picture of the object using QR codes that signal drop off and receipt of the object. Upon completion, this transaction augments player statistics within Swarm! to reflect the success of the transaction. As a result, Eve's public standing within the player community increases, just as it would have if Eve had participated in a coordinated attempt to seize a food source for her colony. Her increased influence within game environment might increase the chances that her next request for a set of AA batteries is also filled.

This mechanic creates an environment in which contributing to the welfare of other Swarm! players through the redistribution of goods and services is rewarded not monetarily, but through the attraction of attention and the generation of influence and repute. The attention attracted by the request is converted into user experience upon completion of the task, allowing the user's behavior to have a more significant impact on the dynamics of the game. Again, this mechanic helps to blur the line between in-game and out-of-game interactions: the in-game world of Swarm! is a distillation and reflection of the everyday out-of-game world of Swarm!'s players. Eve's history as a Swarm! player disposed to help other players in need might be intuitively presented to other members of her colony through features of her trail. When Eve makes a request for aid other players will be more disposed to

respond in kind.[155]

Although our examples have focused on minor transactions of relatively little significance, the game mechanics described here suggest a number of important principles for designing HC systems that harvest the computational dynamics of natural human activity, and the profound impacts these applications might have on a number of vitally important human activities, including education, politics, and urban development. We focus the remaining discussion on economic considerations.

5.4 Naturally optimizing the economy

We can think of the global economy as being a certain kind of HC system in which the problem being computed involves the search for optimal (or near-optimal)[156] allocations of raw materials, labor, and other finite resources ("the economic optimization problem"). This approach to economic theory is broadly called "computational economics" (see e.g. Velupillai, 2000; Norman, 1996), and it takes economic theory to be an application of computability theory and game theory. Historically, some economists have argued that a free capitalist market composed of minimally constrained individual agents (and suitable technological conditions supporting their behavior) provides the most efficient possible economic system (Hayek, 1948). We shall conclude our paper with a discussion of NHC applications as an alternative approach for tackling the economic optimization problem.

Kocherlakota (1998) argues that money is best thought of as a "primitive form of memory" (*ibid.* p. 2). That is, money is a technological innovation that provides a medium for a limited recording of an agent's history of interactions with other agents. On this view, rather than being an intrinsic store of

[155] The influence of perceptions of fairness on economic interactions is an increasingly well-studied phenomenon among economists and psychologists. For a comprehensive overview, see Kolm & Ythier (2006), especially Chapter 8.

[156] The definition of "optimal" is disputed, but the discussion here does not turn on the adoption of a particular interpretation. In general, recall that solving the economic optimization problem involves deciding on a distribution of finite resources (labor, natural resources, &c.). Precisely which distribution counts as "optimal" will depend on the prioritization of values. A robust literature on dealing with conflicting (or even incommensurable) values exists. See, for example, Anderson (1995), Chapter 13 of Raz (1988), and Sen (1997).

value or an independent medium of exchange, money is merely a way to record a set of facts about the past. Kocherlakota argues that this technological role can be subsumed under "memory" in a more general sense, and that while access to money provides opportunities for system behavior that wouldn't exist otherwise, other (more comprehensive) kinds of memory might do all that money does, and more: "...in at least some environments, memory [in the form of high quality information storage and access] may technologically dominate money" (*ibid.* p. 27).

If this characterization is correct, then solving the economic optimization problem involves accomplishing two distinct tasks: identifying precisely *what* information should be recorded in economic memory, and we must devising ways to store and manipulate that information. We might understand Yelp as recording user accounts of a service that attempts to meet these memory challenges. Yelp users leave comments, reviews, and ratings that provide a far more detailed and relevant transaction history with customers than is represented by the relative wealth of the business as a market agent. Luca (2011) finds not only that these reviews have an impact on revenue, but that impact is strengthened with the information content of the reviews, suggesting one place where money may be showing evidence of domination by rich sources of memory.

Swarm! offers a natural approach for meeting the same challenges, in which NHC is leveraged to help solve the economic optimization problem without introducing new economic frictions. This computational work is accomplished through the recording of trails that represents incremental changes in the use history of that location. As Swarm! maps become increasingly detailed and populated they likewise come to function as an effective representation of the attention economy (Simon, 1971; Weng, 2012) in which the saturation of trails around an object approximates a quantitative measure of the value of objects relative to their use[157]. We treat this measure as the aggregate "use-value" of the object (Vargo et al., 2008), and argue that a model of the use-value of objects allows for novel NHC-based

[157] As opposed to value relative to *exchange*. See Marx (1859).

solutions to a variety of standard problems in the optimization of economic systems. A full articulation of the attention economy is not possible here, but we will provide a sketch of one possible implementation using the Swarm! framework.

5.5 Developing the Attention Economy

Recall the central mechanic of Swarm! GPS data about players' movement patterns are aggregated, whether or not a player is actively engaged with the game. Strategically-minded players are rewarded for tagging and modifying the map in a way that gives rise to a detailed reflection of how all Swarm! players use the space covered by the map. The data collected by a Swarm!-like application has the potential to encode many of the facts that might otherwise be encoded less explicitly. Monetary transaction records act as proxy recordings for what we have called *use-value*. The mechanics of Swarm! suggest a way to measure use-value directly by recording how economic agents move through space, how their movement is related to the movement of others, what objects they interact with, the length and circumstances of those interactions, and so on. By tracking this data, we can transform the everyday activities of agents into records of what those agents value and to what degree. This is the "high quality information storage and access" that Kocherlakota suggests may come to technologically dominate currency as economic memory. Still, a number of practical challenges must be surmounted before a NHC/AE based approach to solving the economic optimization problem is realistically viable.

Any implementation of an attention economy in which the economic optimization problem is solved with NHC will clearly involve data collection on a scale that goes far beyond what's possible in Swarm! or with Google Glass, as the mere tracking of gross geospatial position will not record enough information to (for instance) assay the value of individual material objects like pens and light bulbs. Swarm! is an incremental step in that direction, with the more modest and technologically feasible goals of acclimating people to regular engagement with AE platforms, and with developing the social norms appropriate to the demands of an AE. The structure of human communities is strongly coupled to the

technology available during their development. Absent major catastrophes, the sort of ubiquitous computing and social norms necessary for the implementation of an AE will continue to develop in tandem.

Indeed, the success of AE in some sense depends on the development of social customs and attitudes to compensate for the more invasive social coordination technologies that dominated the Industrial Age, which are almost universally characterized by the establishment of hierarchical institutions of control. In such a system, power is concentrated in the hands of the very few, to be executed within very narrow channels of operation. For the disenfranchised, finding ways to circumvent or usurp this power is often a more attractive than accumulating power through so-called "legitimate" means--especially as the powerful increasingly protect their positions through deliberate corruption and abuse, thereby weighting the system heavily against "fair play". In other words, enterprising opportunists looking for success in systems with limited hierarchical control have a disproportionate incentive to "game the system", or exploit loopholes in the rules in ways that give them a disproportionate advantage. Preventing the exploitation of such loopholes requires an ever increasing concentration of power, creating greater incentives to break the system, and greater costs for failing in those attempts. Social customs discouraging such behavior must be imposed from the top, often with violence, as a means of retaining control, since these customs are not reinforced from below.

In contrast, the AE describes a self-organizing system without hierarchical control or concentrations of power, because the rules for operating within the system also support the success of the system as a whole, and so are supported from the bottom without need for top-down enforcement. In other words, the impulse to game an attention economy can be actively encouraged by all parties, since individual attempts to gain a disproportionate advantage within the system simultaneously reinforce the success of the system overall. Recall from section 5.3.3, when Eve snaps a picture of a highly trafficked block. This apparently self-interested act to improve her own in-game resource

collection rate is simultaneously a contribution to the economic optimization problem, and is therefore reinforced by her colony's goals. Of course, Eve is not only rewarded by pursuing self-interested goals; potentially everything Eve does in an attention economy is computationally significant for her community, and therefore her community can support Eve in the pursuit of any goals she wishes without worrying about how her actions might upset the delicate balance of power that supports institutional control. In an attention economy, Eve is not rewarded to the extent that she appeals to existing centers of power; instead, she is rewarded to the extent that her participation has an impact on the development of her community.

We conclude by mentioning some design considerations inspired by Swarm! for building an "Internet of Things" that facilitates the use of NHCs for managing the attention economy. Most obviously, Swarm! is a step toward the creation of pervasive, universally accessible, comprehensive record of the relationship between agents, locations, and objects. As we have said, widespread location and identity tracking of at least *some* sort is essential for the implementation of a true AE. This is a major design challenge in at least two senses: it is a technical engineering challenge, and a social engineering challenge.

The solution to the first challenge will still require technological progress; we do not yet have ubiquitous distribution of the sort of computing devices that would be necessary to implement the fine-grained level of data collection that a real AE would require. In addition to aggregate movement patterns, an AE platform will need to track patterns in the relationships between agents and physical objects. Sterling (2005) introduces the term "spime" to refer to inanimate objects that are trackable in space and time, and broadcast this data throughout their lifetimes. Objects that are designed to live in an attention economy must track more than just their own location and history: they must be able to track their own use conditions, and change state when those use conditions have been met. This will require objects to be sensitive not just to their own internal states, but also to the states of the objects

(and agents) around them: this is the so-called “Internet of Things” (Atzori et al., 2010). There is already some precedent for very primitive functions of this sort. Consider, for instance, the fact that modern high-end televisions often feature embedded optical sensors to detect ambient light levels, and adjust backlighting accordingly for optimal picture quality. We can imagine expanding and improving on that kind of functionality to develop (say) a television that mutes itself when the telephone rings, pauses when you leave the room, or turns itself off when a user engages deeply with another object (for instance a laptop computer) that’s also in the room. These examples are relatively mundane, but they are suggestive of the sort of industrial design creativity and integration needed to design AE-optimized artifacts.

Swarm! approaches this design challenge by imposing some novel clustering methods represented by the caste and colony system. The colony system is a geographical constraint designed to cluster colony members to ensure that they aren’t spread so thin as to undermine the game dynamics. The caste system is a design constraint on the patterns of user activity, and allows users to tell at a glance the functional results of some possible sequence of engagements without knowing too many details about other players. This latter feature is inspired directly by ant colonies, and is important to the organizational dynamics of an AE. In particular, it gives contexts in which it is appropriate for certain agents to have disproportionate influence on some computing task, thereby carving out emergent hierarchies and cliques. The AE/NHC platform is thus applicable to the solution of non-economic social problems, and can be leveraged to help compute solutions to other legal, political, and social puzzles.

As an illustration of how NHCs might be applied to the distribution and management of resources and labor, consider the transaction history for some arbitrary object X. If this record has been reliably maintained on a user-per user basis, it might serve as the basis for resolving disputes about ownership, rights of use, and other coordination problems traditionally settled by legal and political frameworks. If I have years of history driving a specific car on Wednesday mornings, and the use record

shows you driving this car some particular Wednesday morning, then absent some explanation this appears to be a disturbance in use patterns. This information might itself be enough to warrant a complaint through official channels and initiate the machinery of the public justice system to account for this disturbance. In other words, a well-maintained record of the use history of an object might serve as a foundation for NHC solutions to political and legal disputes, and provides a framework for dealing naturally with apparent cases of "stealing" without requiring anything like the disruptive technologies of property, contracts, and other legal frictions.

This is the real heart of the AE/NHC approach to economic optimization: the NHC acts entirely upon data about local patterns of attention, use, and interaction without significantly disturbing the behavioral patterns that generate the data. Rather than indirectly recording facts about my contribution to (or value of) some object or process in monetary memory, which requires its own set of social conventions and techniques to maintain, those facts are recorded *directly* in the history of my relationship to the object or process. We suggest that careful management of those facts, combined with a distributed NHC framework, might allow for a far more efficient economic system than any money-based system.

We've given a characterization of the shape and character of the first of the two design challenges we mentioned above: the technical engineering challenge. While solving this challenge is central to the implementation of the AE, we should not overlook the importance of solving the second challenge either. While technological advances are important, so are advances in the relationship between humans, technology, and society at large. Just as the dissemination of other major, epoch-defining technologies (like the automobile or the telephone) were accompanied by a certain degree of widespread anxiety and social disruption, we expect that the adoption of the ubiquitous computing platforms required for AE implementation (and their concomitant changes in social practice) will be associated with some unrest as society acclimates to some of the underlying changes. In this respect,

Swarm! is more than just an experiment in designing a NHC application--it is an attempt to give society at large a chance to experience the artifacts and socio-cultural practices required for a well-managed AE. The more time we have to grapple with those issues as a community, the smoother the transition to the future will be.

References

Abramson, D. (2008). Turing's responses to two objections. *Minds and Machines*, *18*(2), 147–167.

Ames, M., & Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 971–980). ACM.

Anderson, D. P. (2004). Boinc: A system for public-resource computing and storage. In *Grid Computing, 2004. Proceedings. Fifth IEEE/ACM International Workshop on* (pp. 4–10). IEEE.

Anderson, E. (1995). *Value in ethics and economics*. Harvard University Press.

Aristotle, J. B. (1995). Complete works of Aristotle. *Ed. J. Barnes, Princeton, NJ*.

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer Networks*, *54*(15), 2787–2805.

Auvray, M., Lenay, C., & Stewart, J. (2006). The attribution of intentionality in a simulated environment: the case of minimalist devices. In *Tenth Meeting of the Association for the Scientific Study of Consciousness, Oxford, UK* (pp. 23–26).

Auvray, M., Lenay, C., & Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology*, *27*(1), 32–47.

Baker, L. R. (2004). The ontology of artifacts. *Philosophical Explorations*, *7*(2), 99–111.

Baker, L. R. (2006). On the twofold nature of artefacts. *Studies in History and Philosophy of Science Part A*, *37*(1), 132–136.

Baker, L. R. (2010). Shrinking Difference—Response to Replies. *American Philosophical Association Newsletter on Philosophy and Computers*, *9*(2).

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, *17*(5), 367–386.

Bearman, P. S., Moody, J., & Stovel, K. (2004). Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks1. *American Journal of Sociology*, *110*(1), 44–91.

Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT Press.

Bellomo, N., Delitala, M., & Coscia, V. (2002). On the mathematical theory of vehicular traffic flow I: Fluid dynamic and kinetic modelling. *Mathematical Models and Methods in Applied Sciences*, *12*(12), 1801–1843.

Bertschinger, N., Olbrich, E., Ay, N., & Jost, J. (2008). Autonomy: An information theoretic perspective. *Biosystems*, *91*(2), 331–345.

Bich, L. (2012). Complex emergence and the living organization: An epistemological framework for biology. *Synthese*, *185*(2), 215–232.

Bickhard, M. H. (2007). Mechanism is not enough. *Pragmatics & Cognition*, *15*(3).

Boogerd, F., Bruggeman, F. J., Hofmeyr, J. H. S., & Westerhoff, H. V. (2007). *Systems Biology: Philosophical Foundations*. Elsevier Science.

Borgmann, A. (2009). *Technology and the character of contemporary life: A philosophical inquiry*. University of Chicago Press.

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011). The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference* (pp. 93–102). ACM.

Bosse, T., & Treur, J. (2011). Patterns in world dynamics indicating agency. In *Transactions on computational collective intelligence III* (pp. 128–151). Springer.

Brook, A. (2006). My BlackBerry and Me: Forever One or Just Friends? *Mobile Understanding: The Epistemology of Ubiquitous Communication*, 55–66.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy*, *4*(1), 73–121.

Cash, M. (2010). Extended cognition, personal responsibility, and relational autonomy. *Phenomenology and the Cognitive Sciences*, *9*(4), 645–671.
Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675–684). ACM.
Ciulla, F., Mocanu, D., Baronchelli, A., Gonçalves, B., Perra, N., & Vespignani, A. (2012). Beating the news using social media: the case study of American Idol. *EPJ Data Science*, *1*(1), 1–11.
Clapin, H. (2002). *Philosophy of mental representation*. Oxford University Press.
Clapper, J. R., Young, J. J., Cartwright, J. E., Grimes, J. G., Payton, S. C., Stackley, S., & Popps, D. (2009). FY2009-2034 Unmanned Systems Integrated Roadmap. *Department of Defense: Office of the Secretary of Defense Unmanned Systems Roadmap*.
Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 7–19.
Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT press.
Clark, A. (2001). *Mindware-an Introduction to the Philosophy of Cognitive Science.* Oxford University Press, New York.
Clark, A. (2002a). That special something: Dennett on the making of minds and selves. *Daniel Dennett*, 187–205.
Clark, A. (2002b). "Minds, Brains and Tools" (with a response by Daniel Dennett). In *Hugh Clapin (ed) Philosophy Of Mental Representation*. Oxford: Clarendon Press.
Clark, A. (2004). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.
Clark, A. (2008a). Embodiment and explanation. *Handbook of Cognitive Science. An Embodied Approach*, 41–58.
Clark, A. (2008b). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension: Embodiment, Action, and Cognitive Extension*. Oxford University Press.
Clark, A. (2010). Memento's revenge: The extended mind, extended. *In Richard Menary (ed.), The Extended Mind. Mit Press. 43--66*.
Clayton, P., & Davies, P. (2006). *The re-emergence of emergence*. Oxford University Press Oxford, UK:
Copeland, B. J. (2004). *The Essential Turing*. Oxford University Press.
Corning, P. A. (2002). The re-emergence of "emergence": A venerable concept in search of a theory. *Complexity*, *7*(6), 18–30.
Cummins, R. (1975). Functional analysis. *The Journal of Philosophy*, 741–765.
Cummins, R. (1996). *Representations, targets, and attitudes*. MIT press.
Cummins, R., & Pollock, J. (1995). *Philosophy and AI: Essays at the Interface*. Mit Press.
De La Mettrie, J. O., Bussey, G. C., & II, F. (1912). *Man a machine*. Open court publishing Company.
De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, *6*(4), 485–507.
Dennett, D. C. (1989). *The intentional stance*. MIT press.
Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, *88*(1), 27–51.
Dennett, D. C. (1993). *Consciousness explained*. Penguin UK.
Dennett, D. C. (1998). *Brainchildren: Essays on designing minds*. MIT Press.
Dennett, D. C. (2000). Making Tools for Thinking. In Dan Sperber (Ed.) *Metarepresentations: A Multidisciplinary Perspective (No. 10). Oxford University Press.*
Dennett, D. C. (2001). The zombic hunch: extinction of an intuition? *Royal Institute of Philosophy Supplement*, *48*, 27–43.
Dennett, D. C. (2004). *Freedom evolves*. Penguin UK.
Dennett, D. C. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. MIT Press.
Descartes, R., Weissman, D., & Bluhm, W. T. (1996). *Discourse on the Method: And, Meditations on First Philosophy*. Yale University Press.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification. using game-design elements in non-gaming contexts. In *PART 2—Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (pp. 2425–2428). ACM.

Di Paolo, E., & De Jaegher, H. (2012). The interactive brain hypothesis. *Frontiers in Human Neuroscience*, *6*.

Di Paolo, E. A., Rohde, M., & Iizuka, H. (2008). Sensitivity to social contingency or stability of interaction? Modelling the dynamics of perceptual crossing. *New Ideas in Psychology*, *26*(2), 278–294.

Doniec, A., Mandiau, R., Piechowiak, S., & Espié, S. (2008). A behavioral multi-agent model for road traffic simulation. *Engineering Applications of Artificial Intelligence*, *21*(8), 1443–1454.

Dorigo, M., Bonabeau, E., & Theraulaz, G. (2000). Ant algorithms and stigmergy. *Future Generation Computer Systems*, *16*(8), 851–871.

Dretske, F. I. (1997). *Naturalizing the mind*. mit Press.

Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason*. MIT press.

Ebbs, G. (1996). Can We Take Our Words at Face Value? *Philosophy and Phenomenological Research*, 499–530.

Ebbs, G. (2009). *Truth and Words*. Oxford University Press.

Emmeche, C. (2007). A biosemiotic note on organisms, animals, machines, cyborgs, and the quasi-autonomy of robots. *Pragmatics & Cognition*, *15*(3), 455–483.

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the Economics of Giving, Altruism and Reciprocity*, *1*, 615–691.

Ferber, J., & Gutknecht, O. (1998). A meta-model for the analysis and design of organizations in multi-agent systems. In *Multi Agent Systems, 1998. Proceedings. International Conference on* (pp. 128–135). IEEE.

Ferber, J., Gutknecht, O., & Michel, F. (2004). From agents to organizations: an organizational view of multi-agent systems. In *Agent-Oriented Software Engineering IV* (pp. 214–230). Springer.

Fortunato, S., & Castellano, C. (2012). Community structure in graphs. In *Computational Complexity* (pp. 490–512). Springer.

Franklin, S. (2003). A conscious artifact? *Journal of Consciousness Studies*, *10*(4-5), 4–5.

Franklin, S., & Graesser, A. (1997). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *Intelligent agents III agent theories, architectures, and languages* (pp. 21–35). Springer.

Franssen, M. (2006). The normativity of artefacts. *Studies in History and Philosophy of Science Part A*, *37*(1), 42–57.

French, R. M. (2000). The Turing Test: the first 50 years. *Trends in Cognitive Sciences*, *4*(3), 115–122.

Froese, T., Iizuka, H., & Ikegami, T. (2014). Embodied social interaction constitutes social cognition in pairs of humans: A minimalist virtual reality experiment. *Scientific Reports*, *4*.

Gallagher, S., Hutto, D. D., Slaby, J., & Cole, J. (2013). The brain as part of an enactive system. *Behavioral and Brain Sciences*, *36*(04), 421–422.

Gehlen, A., Lipscomb, P., & Berger, P. L. (1980). *Man in the Age of Technology*. Columbia University Press New York.

Gordon, D. M. (2010). *Ant encounters: interaction networks and colony behavior*. Princeton University Press.

Greene, K., Thomsen, D., & Michelucci, P. (2012). Massively collaborative problem solving: new security solutions and new security risks. *Security Informatics*, *1*(1), 1–17.

Greene, K. A., & Young, T. A. (2013). Building blocks for collective problem solving. In *Handbook of Human Computation* (pp. 347–365). Springer.

Harnad, S. (1992). The Turing Test is not a trick: Turing indistinguishability is a scientific criterion. *ACM SIGART Bulletin*, *3*(4), 9–10.

Haraway, D. (1991). A cyborg manifesto: science, technology, and socialist-feminism in the late twentieth century. *Simians, Cyborgs and Women: The Reinvention of Nature*, 149–82.
Haselager, W. F. (2005). Robotics, philosophy and the problems of autonomy. *Pragmatics & Cognition*, *13*(3).
Haselager, W. F., & Gonzalez, M. E. Q. (2007). Mechanicism and autonomy: What can robotics teach us about human cognition and action? *Pragmatics & Cognition*, *15*(3), 407–412.
Haugeland, J. (1981). *Mind Design: Philosophy, Psychology, Artificial Intelligence*. Montgomery, Vt.: Bradford Books.
Haugeland, J. (1997). *Mind design II: philosophy, psychology, artificial intelligence*. MIT press.
Haugeland, J. (2002). Authentic intentionality. *In Scheutz, M. (Ed.). Computationalism: New Directions. The MIT Press.*
Hayek, F. A. (1948). *Individualism and economic order*. University of Chicago Press.
Heersmink, R. (2014). The metaphysics of cognitive artefacts. *Philosophical Explorations*, (ahead-of-print), 1–16.
Heidegger, M. (1954). The question concerning technology. *Technology and Values: Essential Readings*, 99–113.
Hodson, H. (2012). Google's Ingress game is a gold mine for augmented reality. *New Scientist*, *216*(2893), 19.
Houkes, W., & Meijers, A. (2006). The ontology of artefacts: the hard problem. *Studies In History and Philosophy of Science Part A*, *37*(1), 118–131.
Hwang, T., Pearce, I., & Nanis, M. (2012). Socialbots: Voices from the fronts. *Interactions*, *19*(2), 38–45.
Jackson, F., & Pettit, P. (1990). Program explanation: A general perspective. Analysis, 107–117.
Johnson, N., Zhao, G., Hunsader, E., Qi, H., Johnson, N., Meng, J., & Tivnan, B. (2013). Abrupt rise of new machine ecology beyond human response time. *Scientific Reports*, *3*.
Kauffman, S. A. (2002). *Investigations*. Oxford University Press.
Keller, E. F. (2008). Organisms, Machines, and Thunderstorms: A History of Self-Organization, Part One. *Historical Studies in the Natural Sciences*, *38*, 45–75.
Keller, E. F. (2009). Organisms, Machines, and Thunderstorms: A History of Self-Organization, Part Two. Complexity, Emergence, and Stable Attractors. *Historical Studies in the Natural Sciences*, *39*, 1–31.
Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popović, Z., … Players, F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, *108*(47), 18949–18953.
Kline, S. J. (1985). What is technology? *Bulletin of Science, Technology & Society*, *5*(3), 215–218.
Kocherlakota, N. R. (1998). Money is memory. *Journal of Economic Theory*, *81*(2), 232–251.
Kolm, S.-C., & Ythier, J. M. (2006). Handbook of the economics of giving, altruism and reciprocity: Foundations (Vol. 1). Elsevier.
Kroes, P. (2012). Theories of technical functions. In *Technical Artefacts: Creations of Mind and Matter* (pp. 47–88). Springer.
Kroes, P., & Meijers, A. (2006b). The dual nature of technical artefacts. *Studies in History and Philosophy of Science Part A*, *37*(1), 1–4.
Ladyman, J., Ross, D., Spurrett, D., & Collier, J. G. (2007). *Every Thing Must Go: Metaphysics Naturalized*. Clarendon Press.
Ladyman, J. (2007). Scientific structuralism: On the identity and diversity of objects in a structure. In *Aristotelian Society Supplementary Volume* (Vol. 81, pp. 23–43). Wiley Online Library.
Latour, B. (1999). *Pandora's hope: essays on the reality of science studies*. Harvard University Press.
Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012). Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web* (pp. 251–260). ACM.

Leopold, A. (1949). A Sand County Almanac (Vol. 1). New York: Oxford University Press.
Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science*, *80*(2), 241–261.
Lovelace, A. K. C. of, & Menabrea, L. F. (1843). *Sketch of the analytical engine invented by Charles Babbage, Esq*. Richard and John E. Taylor.
Luca, M. (2011). *Reviews, reputation, and revenue: The case of Yelp. com*. Harvard Business School.
Marx, K., & Dobb, Maurice (ed). (1979). *A contribution to the critique of political economy*. New York: International Publishers.
McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world*. Penguin.
McLaughlin, P. (2001). *What functions explain: Functional explanation and self-reproducing systems*. Cambridge University Press.
McLaughlin, P. (2002). On Having a Function and Having a Good. *Analyse & Kritik*, *24*, 130–43.
Menary, R. (2010). *The extended mind*. MIT Press.
Meyer, J.-A., Husbands, P., & Harvey, I. (1998). Evolutionary robotics: A survey of applications and problems. In *Evolutionary Robotics* (pp. 1–21). Springer.
Michelucci, P. (2013). *Handbook of human computation*. Springer.
Mira, J. M. (2008). Symbols versus connections: 50 years of artificial intelligence. *Neurocomputing*, *71*(4), 671–680.
Mitri, S., Floreano, D., & Keller, L. (2009). The evolution of information suppression in communicating robots with conflicting interests. *Proceedings of the National Academy of Sciences*, *106*(37), 15786–15790.
Mitter, S., Wagner, C., & Strohmaier, M. (2014). Understanding the impact of socialbot attacks in online social networks. *arXiv Preprint arXiv:1402.6289*.
Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *Intelligent Systems, IEEE*, *21*(4), 18–21.
Morrison, F. (2008). *The Art of Modeling Dynamic Systems: Forecasting for Chaos, Randomness, and Determinism*. Courier Dover Publications.
Muntean, I., & Wright, C. D. (2007). Autonomous agency, AI, and allostasis: A biomimetic perspective. *Pragmatics & Cognition*, *15*(3).
Norman, A. L. (1994). Computability, complexity and economics. *Computational Economics*, *7*(1), 1–21.
Omicini, A., Ricci, A., Viroli, M., Castelfranchi, C., & Tummolini, L. (2004). Coordination artifacts: Environment-based coordination for intelligent agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1* (pp. 286–293). IEEE Computer Society.
Paruchuri, P., Pullalarevu, A. R., & Karlapalem, K. (2002). Multi agent simulation of unorganized traffic. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1* (pp. 176–183). ACM.
Pavlic, T. P., & Pratt, S. C. (2013). Superorganismic behavior via human computation. In *Handbook of Human Computation* (pp. 911–960). Springer.
Prem, E. (1997). Epistemic autonomy in models of living systems. In *Proceedings of the fourth European conference on artificial life* (pp. 2–9).
Preston, B. (2009). Philosophical theories of artifact function. *Philosophy of Technology and Engineering Sciences*, *9*, 213–233.
Putnam, H. (1975). The meaning of "meaning." *In A. Pessin & S. Goldberg (Eds.), The Twin Earth Chronicles. London: M. E. Sharpe.*
Quine, W. (1951). Two Dogmas of Empiricism. *The Philosophical Review*, 20–43.

Rashid, A. M., Ling, K., Tassone, R. D., Resnick, P., Kraut, R., & Riedl, J. (2006). Motivating participation by displaying the value of contribution. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 955–958). ACM.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web* (pp. 249–252). ACM.

Raz, J. (1986). *The morality of freedom*. Oxford University Press.

Ridder, J. de. (2006). Mechanistic artefact explanation. *Studies in History and Philosophy of Science Part A*, *37*(1), 81–96.

Rohde, M. (2010). *Enaction, embodiment, evolutionary robotics: simulation models for a post-cognitivist science of mind* (Vol. 1). Springer.

Rohde, M., & Stewart, J. (2008). Ascriptional and "genuine"autonomy. *BioSystems*, *91*(2), 424–433.

Salvini, P., Laschi, C., & Dario, P. (2010). Design for acceptability: improving robots' coexistence in human society. *International Journal of Social Robotics*, *2*(4), 451–460.

Saygin, A. P., Cicekli, I., & Akman, V. (2003). Turing test: 50 years later. In *The Turing Test* (pp. 23–78). Springer.

Scheele, M. (2006). Function and use of technical artefacts: social conditions of function ascription. *Studies In History and Philosophy of Science Part A*, *37*(1), 23–36.

Scheutz, M. (2002). *Computationalism: New Directions*. MIT Press.

Schilbach, L. (2010). A second-person approach to other minds. *Nature Reviews Neuroscience*, *11*(6), 449–449.

Schreibman, S., Siemens, R., & Unsworth, J. (2008). *A companion to digital humanities*. John Wiley & Sons.

Schyfter, P. (2009). The bootstrapped artefact: A collectivist account of technological ontology, functions, and normativity. *Studies in History and Philosophy of Science Part A*, *40*(1), 102–111.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(03), 417–424.

Searle, J. R. (1984). *Minds, brains, and science*. Harvard University Press.

Sellars, W. (1997). *Empiricism and the Philosophy of Mind*. Harvard University Press.

Sen, A. (1997). Maximization and the Act of Choice. *Econometrica: Journal of the Econometric Society*, 745–779.

Sharkey, N. (2008). The ethical frontiers of robotics. *Science*, *322*(5909), 1800–1801.

Shim, H. (2007). Establishing a Korean robot ethics charter. In *IEEE ICRA workshop on roboethics, April* (Vol. 14, p. 2007).

Simon, H. A. (1971). Designing organizations for an information-rich world. *Computers, Communications, and the Public Interest*, *72*, 37.

Simon, H. A. (1996). *The sciences of the artificial*. MIT press.

Smith, A. (2012). The best (and worst) of mobile connectivity. *Pew Internet & American Life Project*. Retrieved from http://www.pewinternet.org/fact-sheets/mobile-technology-fact-sheet/

Sribunruangrit, N., Marque, C. K., Lenay, C., Hanneton, S., Gapenne, O., & Vanhoutte, C. (2004). Speed-accuracy tradeoff during performance of a tracking task without visual feedback. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, *12*(1), 131–139.

Sterling, B., Wild, L., & Lunenfeld, P. (2005). *Shaping things*. MIT press Cambridge, MA.

Strogatz, S. H. (2001). Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering (studies in nonlinearity).

Stuart, S. (2002). A Radical Notion of Embeddedness A Logically Necessary Precondition for Agency and Self-Awareness. *Metaphilosophy*, *33*(1-2), 98–109.

Sugiyama, Y., Fukui, M., Kikuchi, M., Hasebe, K., Nakayama, A., Nishinari, K., … Yukawa, S. (2008). Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam. *New Journal of Physics*, *10*(3), 033001.

Susi, T., & Ziemke, T. (2001). Social cognition, artefacts, and stigmergy: A comparative analysis of theoretical frameworks for the understanding of artefact-mediated collaborative activity. *Cognitive Systems Research*, *2*(4), 273–290.

Tani, J. (2009). Autonomy of Self at criticality: The perspective from synthetic neuro-robotics. *Adaptive Behavior*, *17*(5), 421–443.

Taylor, M. A., Woolley, J. E., & Zito, R. (2000). Integration of the global positioning system and geographical information systems for traffic congestion studies. *Transportation Research Part C: Emerging Technologies*, *8*(1), 257–285.

Thomasson, A. (2007). Artifacts and human concepts. *Creations of the Mind: Theories of Artifacts and Their Representation*, 52–73.

Toepfer, G. (2012). Teleology and its constitutive role for biology as the science of organized systems in nature. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 113–119.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433.

Turing, A. (2004). Lecture on the Automatic Computing Engine (1947). *In Copeland, B. J. (Ed.). (2004). The Essential Turing. Oxford University Press.*, 362.

Vaccari, A. (2013). Artifact Dualism, Materiality, and the Hard Problem of Ontology: Some Critical Remarks on the Dual Nature of Technical Artifacts Program. *Philosophy & Technology*, *26*(1), 7–29.

Vaesen, K. (2011). The functional bias of the dual nature of technical artefacts program. *Studies in History and Philosophy of Science Part A*, *42*(1), 190–197.

Vargo, S. L., Maglio, P. P., & Akaka, M. A. (2008). On value and value co-creation: A service systems and service logic perspective. *European Management Journal*, *26*(3), 145–152.

Velupillai, K. (2000). *Computable economics: the Arne Ryde memorial lectures* (Vol. 5). Oxford University Press.

Vermaas, P. E., Carrara, M., Borgo, S., & Garbacz, P. (2013). The design stance and its artefacts. *Synthese*, *190*(6), 1131–1152.

Veruggio, G. (2006). The euron roboethics roadmap. In *Humanoid Robots, 2006 6th IEEE-RAS International Conference on* (pp. 612–617). IEEE.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, *325*(5939), 425.

Von Ahn, L. (2009). Human computation. In *Design Automation Conference, 2009. DAC'09. 46th ACM/IEEE* (pp. 418–419). IEEE.

Von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*(8), 58–67.

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. *Science*, *321*(5895), 1465–1468.

Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*, *2*.

Wilde, O. (1909). The Soul of Man. Boston: J.W. Luce & Co.