Trust and multi-agent systems: applying the "diffuse, default model" of trust to experiments involving artificial agents

Jeff Buechner · Herman T. Tavani

Published online: 9 September 2010

Abstract We argue that the notion of trust, as it figures in an ethical context, can be illuminated by examining research in artificial intelligence on multi-agent systems in which commitment and trust are modeled. We begin with an analysis of a philosophical model of trust based on Richard Holton's interpretation of P. F. Strawson's writings on freedom and resentment, and we show why this account of trust is difficult to extend to artificial agents (AAs) as well as to other non-human entities. We then examine Margaret Urban Walker's notions of "default trust" and "default, diffuse trust" to see how these concepts can inform our analysis of trust in the context of AAs. In the final section, we show how ethicists can improve their understanding of important features in the trust relationship by examining data resulting from a classic experiment involving AAs.

Keywords Artificial agents · Default trust · Diffuse trust · Multi-agent systems · Trust

Abbreviations

AAs	Artificial agents
DSCP	Discount-based social-commitment policy
RSCP	Ranking-based social-commitment policy
SPIRE	Shared plans intention reconciliation experiments

Introduction

In the past decade, fairly extensive literatures have emerged on two ICT-ethics related topics: artificial agents (AAs) and e-trust (i.e., trust in the context of cyberspace or digital environments). An examination of the literature on AAs suggests a focus on issues that range from questions about whether AAs can qualify as rational, autonomous, and (fully) moral agents to questions about whether (and, if so, to what extent) AAs can be held accountable for their actions. And questions pertaining to e-trust have generally focused on whether individuals can trust: (a) the Web sites with which they interact (including the transactions they make on e-commerce sites, as well as the integrity of information available to them on many non-commercial sites), and (b) other individuals with whom they interact online (especially in purely virtual contexts where the individuals involved have never met one another in physical space). For the most part, there has been little overlap in ICT-ethics-related research on these two topics, which

This essay is the first in a series of papers in which we examine the role of trust in multi-agent systems. In subsequent papers we consider the issue of whether artificial agents can trust and be trusted, the connection between frame problems in artificial intelligence and the trust relation, the conceptual connection between the moral notion of trust and the epistemic notion of trust, and the complex network of connections between artificial agents and human agents with respect to both moral trust and epistemic trust.

J. Buechner
Department of Philosophy, Rutgers University, Newark,
NJ, USA
e-mail: buechner.rci@rutgers.edu

J. Buechner
Saul Kripke Center, City University of New York-The Graduate
Center, New York, USA
e-mail: JBuechner@gc.cuny.edu

H. T. Tavani (✉)
Department of Philosophy, Rivier College, 420 Main St.,
Nashua, NH 03060, USA
e-mail: htavani@rivier.edu

might be interpreted to suggest that key questions and controversies affecting AAs and trust are not related in any philosophically interesting sense. In the past few years, however, questions about some important connections between these distinct categories of ICT-ethics research have begun to be examined.[1]

In the present essay, we seek to gain a better understanding of both the concept of trust and the role that it plays in the context of AAs. However, we also note at the outset that there are many open questions affecting both notions that cannot be examined here. In the case of AAs, for example, critical philosophical questions arise that affect concepts such as agency, autonomy, rationality, moral responsibility, and so forth.[2] Because we cannot address these issues in the detail they warrant, we proceed on the basis of what some might view as a minimalist or "thin" notion of an artificial agency—i.e., one that does not presuppose a robust account of AAs as moral agents or (necessarily) as agents that are fully autonomous or even fully rational. We do not believe that we have to defend the latter kind of account of AAs to support the claims we that make in this essay pertaining to what we call "multi-agent systems."[3]

With respect to the concept of trust, we also note that there are many open questions that cannot be examined in the present essay. There are also some important conceptual connections between trust and related notions that cannot be analyzed within the scope of this essay. These include questions about trust as it relates to security (Nissenbaum 2001), privacy (Fried 1990), autonomy (O'Neill 2002), friendship (Weckert 2005), and risk, especially in the context of e-commerce (Camp 2000). Additionally, there are some important conceptual and practical distinctions affecting trust that cannot be addressed in the detail they deserve—e.g., questions involving ethical vs. epistemic aspects of trust (Simon 2009), and questions about whether trust can be viewed as "affective and conative" as well as cognitive (Baier 1986). These questions, as in the case of many of the current controversial questions pertaining to AAs that fall outside the scope of this essay, are not crucial to our analysis of trust; so, there is no need to review the vast literature on trust here.[4]

We argue that our understanding of the notion of trust, especially as it figures in an ethical context, can be illuminated by examining work in artificial intelligence (AI) on multi-agent systems in which commitment and trust are modeled.[5] We believe that ethicists should make use of relevant data from the experiments conducted in multi-agent systems, because it can be used to test ethical hypotheses about both the notion of trust and the structure of ethical theories of trust. Furthermore, conducting these experiments would not embroil the ethicist in the kinds of moral problems involved with using human subjects for experiments. (This is a feature that is either little known or not at all known to ethicists and to AI theorists conducting experiments with complex multi-agent software systems.) The remaining sections of this essay are organized into two main parts. In Part I, we present a brief overview of the concept of trust and we defend Margaret Urban Walker's notions of "default trust" and "diffuse trust" as ideal models for analyzing issues affecting trust in multi-agent systems. Part II aims at showing why ethicists can benefit from the research on trust in experiments involving artificial agents.

Part I: trust and artificial agents

The notion of "trust" figures importantly in the moral relations between people, animals, and institutions, as well as the ways in which we regard our environment. However,

[1] See, for example, Taddeo (2008); Lim et al. (2008) and Grodzinsky et al. (2009).

[2] In a provocative (and now, arguably, classic) paper, Floridi and Sanders (2004) suggest that AAs can be fully rational, autonomous, and moral agents (in addition to being "moral patients"). They also contrast autonomous artificial agents with what they call "heteronomous artificial agents." But not everyone has accepted the view that AA's can qualify as moral agents. For example, Johnson (2006) argues that because AAs lack intentionality, they cannot be viewed as moral agents, even if they satisfy the requirements for "moral entities." And Himma (2009) argues that for AAs to qualify as moral agents, they will have to convincingly demonstrate that they posses a range of properties, including consciousness. Others have questioned whether AAs can be "responsible entities" and some suggest that we may need to expand the notion of responsibility because of the kinds of accountability issues raised by AAs (Stahl 2006). However, we will not pursue these controversies here. Basically, we agree with Moor (2006) that there are at least four distinct levels in which AAs can have an "ethical impact," even if we cannot attribute moral accountability to AAs themselves and even if AAs do not qualify as (full) moral agents.

[3] We use the standard definition of an artificial agent introduced by Subrahamanian, et al. (2000, p. 4), which easily enables us to define a multi-agent system as well. On this definition, artificial software agents consist of a collection of software providing useful services that other agents might employ, a description of the services accessible to other agents, the capacity for autonomous actions, the capacity to describe how that agent determines which specific courses-of-action to take, and the capacity to interact with other agents, both artificial and human, in different environments, ranging from cooperative to hostile.

[4] For some excellent discussions and analyses of classic definitions and formulations of trust, see Luhmann (1979); Gambetta (1998), and Taddeo (2009).

[5] We also believe that researchers working on multi-agent systems who model the notions of commitment and trust in their experiments can benefit from examining the philosophical literature on trust. However, this argument is developed in a separate paper.

the minimal set of necessary and sufficient conditions found in a dictionary definition of "trust," does little to provide assistance in elucidating the nature of the *moral component* in trust. Indeed, the primary definition of trust as "firm reliance on the integrity, ability, or character of a person or thing,"[6] fails to settle what the nature of the dependency consists in, and how "reliance on ability" can differ from reliance on either integrity or character. The philosophical literature on trust aims to remedy this defect by providing an explication of the notion that situates its moral focus primarily in the relations between persons, and it describes conditions under which it applies to both institutional and environmental—especially the environment of a community—relations.

Trust, normative expectations, and responsibility

Often we harbor expectations about how people will behave in certain contexts, where our expectations are based on both observations of their behavior and predictions about how they will behave in the future. We expect an agent, Jones, to behave in a certain way when we encounter him, in part because we have observed this behavior in the past, and in part because we predict that he will continue to engage in this behavior in the future. But such expectations are hardly what make the relationship between *A* and *B* a matter of trust on *A's* part that *B* do such-and-such in a given context. One can expect someone to do such-and-such, but not necessarily trust them to do it. Moreover, one can trust someone to do such-and-such, though not expect them to do it. (I trust Shimura to deliver the legal papers to me, but since he has been ill recently, I do not expect that he will deliver them to me tonight.) Thus, a non-normative notion of expectation is neither necessary nor sufficient for trust.

There is something more that needs to be added to this description in order to arrive at the moral notion of trust. In particular, what must be added is a normative component, because without it any notion of expectation about what others will do would hardly be said to have a moral bearing. One of the most well-known additions has a complex history whose origin can be traced to P. F. Strawson.[7]

Strawson characterized certain participant-reactive attitudes, such as resentment and gratitude, in terms of two normative components: excuses and justifications. Richard Holton applied the Strawsonian conception of reactive attitudes to the trust relation, suggesting that to trust someone is to rely on them to do something and to *react* either (a) with resentment to their not doing it or (b) with gratitude for their doing it. In this scheme, both resentment and gratitude can be modified by either excuses or justifications.[8]

Margaret Urban Walker further explicates the normative components in the trust relation by modifying Holton's account somewhat; specifically, she introduces the notion of a normative expectation on the part of the person who entrusts another, which brings with it a requirement of responsibility on the part of the entrusted. Combining the views of Strawson, Holton, and Walker, we believe that a relationship of trust between *A* and *B* is one in which the following five conditions obtain:

(i) *A* has a normative expectation (which may be based on a reason or motive) that *B* will do such-and-such;
(ii) *B* is responsible for what it is that *A* normatively expects her to do;
(iii) *A* has the disposition to normatively expect that *B* will do such and such responsibly;
(iv) *A's* normative expectation that *B* will do such-and-such can be mistaken;
(v) [Subsequent to the satisfaction of conditions (i)–(iv)] *A* develops a disposition to trust *B*.

A normative expectation is in place when "I rely on you to do what you should. I do not only expect *that* you will do it, I expect it *of* you."[9] *A* can normatively expect that *B* will do such-and-such, even when *A* does not expect that *B* will do it. That is, *A* does not predict that *B* will do it, but still expects it of him. Because the normative components of the trust relation have their origin in Strawson, we call this view of trust the Strawson-Holton-Walker conception of trust.

That *A* has the disposition to trust *B* to do such-and-such allows room for cases in which *B* does not do such-and-such, even though *A* trusts that *B* will do it. Condition (iv) allows for cases in which *A* mistakenly trusts *B*. There may be over-determined cases in which *A* expects *B* to do such-and-such both because *A* trusts *B* and because *A* has been coerced into expecting *B* to do such-and-such. As long as *A* normatively expects *B* to do such-and-such, and it is not the case that the normative expectation is itself the result of being coerced, there is a genuine trust relationship in place

[6] *American Heritage College Dictionary* (Fourth edition), Houghton Mifflin Company, 2002.

[7] Peter Strawson, "Freedom and Resentment," in *Freedom and Resentment and Other Essays*, Routledge, 1974, pp. 1–28. Strawson introduces his distinction between the "participant attitude toward others" and an "objective attitude towards others," where the participant attitude brings in the normative notion of responsibility to others, while the objective attitude is characterized by non-normative, descriptive features.

[8] See Richard Holton, "Deciding to Trust, Coming to Believe," in *Australasian Journal of Philosophy*, **72** (1994), pp. 63–76.

[9] Margaret Urban Walker, *Moral Repair: Reconstructing Moral Relations After Wrongdoing*, Cambridge University Press, 2006, p. 79.

(provided the other conditions are satisfied as well). Finally, the definition rules out as instances of trust for those cases in which *B* has a responsibility to do such-and-such, even though *A* does not attribute responsibility to do such-and-such to *B*.

The notions of normative expectation and of responsibility, then, are what the normative features of trust consist in; these characteristics, in turn, make it difficult to extend the Strawson-Holton-Walker account to relationships between persons and non-human entities such as institutions, animals, and artificial agents (including multi-agent software systems). The kind of normative expectations that we, as humans, have toward other people appear to differ from the kinds of expectations we have toward multi-agent systems in a morally important way—namely, we can hold other people responsible for what it is that we trust them to do, but we cannot hold multi-agent systems responsible for what it is that we rely upon them to do. For one thing, current multi-agent systems are not completely free to do what they do; rather, they are programmed (at least to some extent) to do what they do. In that sense, they would seem compelled to do such-and-such. We can, of course, hold the designer of the multi-agent system (or the group of designers, or even the institution in which the software team works) responsible for their product.

Is there room in the moral notion of trust to accommodate normative relations between persons and the multi-agent systems with which they interact? Of course, one might object to either the responsibility requirement or the normative-expectation requirement for the notion of trust; however, without these requirements, we are left with an impoverished notion of trust that it is difficult to apply and thus not very useful. A virtue of the Strawson-Holton-Walker conception of trust that it is able to capture most of the considerations, examples, and distinctions that fall under what we believe to be a robust notion of trust.

Default trust and zones of default trust

How can the notion of trust apply in the case of AAs other than in the context in which the human creator of those agents has a trusting relation with all those humans who interact with the agents? Could there be conditions under which the AAs are a functional part of an environment in which individual humans enter into trusting relationships with that environment—where there are other individual humans in that environment? In an excellent analysis of trust (in connection with her account of the concept of "moral repair"), Margaret Urban Walker proposes her notion of default trust:

> Sometimes when people refer to their 'communities,' either as networks of people or as geographical locations or both, they capture this sense of the place where one feels relatively safe. This is not because one believes one is utterly protected, but because one believes one knows what to expect and from whom to expect it, and one knows what is normal and what is out of place. One knows, in a word, what to expect and whom to trust. This practical outlook of ease, comfort, or complacency that relies on the good or tolerable behavior of others is the form of trust I call *default trust.*[10]

Walker allows for "zones" of default trust—these are the "spaces" and circumstances in which default trust is experienced in any degree and in whatever way. The importance of allowing default trust within a default zone to occur in any degree and in whatever way is that we can now speak of cases of various kinds that are intrinsically different, but whose common feature is that they involve a zone of default trust. In that way, the concept of a zone of default trust can do much of the work in assimilating a wide range of disparate cases. How, though, can the notion of an artificial agent have a conceptual footing in the notion of the zone of default trust? To see how AAs can be accommodated within zones of default trust in such a way that we would take them to be responsible agents, we need to examine some examples of default trust.

The examples that Walker employs mainly have to do with communities in which large numbers of individuals live and work together. For instance, a city is a place where we rely upon other individuals to follow procedures, rules, and modes of conduct that will not place us in jeopardy in everyday situations, such as crossing the street or in driving one's car on the roads. We expect—and this expectation is normative—that another individual living in the city will drive on the correct side of the street. Moreover, we take this to be a responsibility that person has toward others in her community. Although there are laws that require this to be the appropriate conduct in driving one's car, there are also cases where there are no laws governing safe conduct. For instance, in walking on a city street, one expects not to walk into a person walking in the opposite direction. Moreover, this is a normative expectation, and it is the responsibility of the other person not to walk into you. Notice, as well, that this example is one in which the requirements are symmetric and reversible—that is, *A* trusts *B* and *B* trusts *A*. In other words, *A* normatively expects *B* to do such-and-such and *B* is responsible for doing it, and *B* normatively expects *A* to do the same, and *A* is responsible for doing it as well.

Walker describes a zone of default trust as an "unreflective and habitual background … for specific episodes

[10] Walker, *op. cit.,* p. 85.

and relationships of trust and entrusting."[11] That is, the zone of default trust is one in which people engage in habitual behavior of certain kinds, and in which one does not consciously reflect on the kinds of behavior that are appropriate. Rather, one simply engages in that behavior, with little or no conscious reflection. Consider the following scenario:

> You are connected to a community of people via the internet. There are various kinds of transactions and behaviors that you can engage in; suppose that one of them is an auction of some sort, such as eBay. There are many (human) individuals involved in this environment, as well as many different kinds of artificial agents. Certainly, it is the case that a participant in the auction does not know who the technical support people are, who the financial advisers are, who the creators of the auction website are, and so forth. Nonetheless, there are relations of trust between various individuals in this environment; as in the case of a city, or a hospital, or a department store, the electronic auction site, too, is a zone of default trust.

One feature of a zone of default trust that needs articulation is that there are many individuals with whom one will never have any sort of personal contact, and yet one has a trusting relationship with them, and they have with the individual. That is an important feature of default trust: the individuals with whom one has a trusting relationship need not be specified in advance, and need not ever engage in any kind of behavior that affects those who trust them. Similarly, the kinds of behaviors in which individuals in a zone of default trust engage might personally affect another individual, even though both individuals never have direct contact with one another, and even though neither individual will ever know each other—that is, *A* will not know that *B* is the locus of the behavior that affects *A*, and *B* will not know that her behavior affects *A*. (It is important to see that our definition of trust does not exclude such cases.)

Thus, in addition to the real individuals who occupy zones of default trust, there are also both generic individuals and objects that are surrogates for individuals. So long as the expectations one has about what these generic individuals and surrogates for individuals will have a normative component, and so long as the generic individuals and surrogates are responsible for what they do, the relationship is one of trust and not merely a reliance that has no normative component. The extension of the notion of trust from individuals to generic individuals, and from individuals to surrogates for individuals, is intrinsic to the notion of a zone of default trust.

It might be objected that moving the locus of trust away from actual individuals to generic individuals and surrogates for individuals is illegitimate because such constructions do not have responsibility for what is done in a given context. One might also object to the view that we cannot have normative expectations of what these entities will do, since they are not actual individuals toward whom we can and do have normative expectations and they are not entities that can act either responsibly or irresponsibly. In other words, one cannot get normativity out of questionable constructions such as generic individuals and surrogates without providing additional arguments. Although Walker does not consider this objection (since she does not introduce generic individuals and surrogates into her discussion), we believe that it needs to be addressed. Suppose, for example, that one engages in a trusting relation with another individual. It is perfectly natural to query, counterfactually, what that individual would do in such-and-such counterfactual circumstances. We do not say of such queries that they are illegitimate because either (a) one cannot have normative expectations of what does not exist, or (b) non-existent individuals cannot act in ways that are responsible or irresponsible.

We should note that such an objection to considering counterfactuals of this sort would have little force, since counterfactual considerations of moral situations are part-and-parcel of standard episodes of moral reasoning and of the engagement of the moral imagination. We often counterfactually imagine what is the right course of action to perform, even though the situation that we imagine does not (and might never) exist and the counterfactual individuals populating that imagined situation do not (and might never) exist. A common example of such uses of the moral imagination is in writing fiction that has moral lessons—such as the novels of Iris Murdoch. We would sound foolish if we said that it is not legitimate to take seriously the moral lessons of her novel, *The Black Prince*, since the characters in it do not exist and the situations she describes in it do not exist.[12]

Diffuse default trust

Walker also introduces an extension of her notion of default trust, which she calls "diffuse default trust."[13] To illustrate this concept, Walker employs an example in which one encounters particularly bad service on an airplane operated by a major commercial airline. When we have such an experience of bad service, it is appropriate to

[11] Ibid., p. 85.

[12] See, for example, Murdoch (1973). We cannot argue for the legitimacy of our claim here, since doing so would require additional space for arguments to show how the metaphysics of fictional entities allows a natural explanation to be provided of how we can have normative relations to fictional entities.

[13] Walker, *op. cit.*, pp. 85ff.

feel resentment, not necessarily toward specific individuals who work for the airline, but toward the airline itself. This is an important point: the feeling of resentment reveals that there is a normative component in the expectations one had about the kind of service that should have been provided by the airline. Of course, one does not feel resentment toward the weather when one's expectations of it have been defeated, since one does not enter into relations of trust with the weather. On the contrary, it would be foolish to say to someone that they should not resent the bad service of an airline, because you cannot resent the airline, but only the individuals who work for the airline.

It is foolish, because we can and do feel such resentment. It is a datum that needs explanation in the context of developing an adequate account of trust, and it is the virtue of Walker's account that she provides conceptual space for it. She says, of the bad service of the airline, that

> … the lapses are not directed at us, nor do we necessarily know exactly whom we blame … We might say that in this, as in many cases of even more diffuse default trust, we did not rely on X to do A, and Y to do B, and so on (although we might have some beliefs of this kind), but rather that we expected 'reliable, courteous, and orderly service' of the airline, which is not just a particular group of unnamed individuals but a mode of organization that is supposed to train and enable whatever individuals are filling organizational roles to perform effectively to the end we rely on.[14]

Of course, it need not be the case that our reliance on such institutional entities in a relation of diffuse default trust is realistic, but that does not show that the relation is not one of trust at all. Rather, it simply shows that we can have unrealistic normative expectations.

In making room for artificial agents in our account of trust via the employment of notions such as default trust, zones of default trust, and diffuse default trust, we are now in a position to develop a key claim of this paper—viz., that ethicists can learn about central features of the trust relation by examining experiments involving artificial agents (that do not involve any human individuals). One might assume, initially at least, that a study of artificial agents in large complex software systems would have no implications for issues affecting trust that are of interest to ethicists. However, we intend to show that this view is both naïve and false.

Artificial agents

Artificial agents have been a staple of computer science for many decades. They are modeled on the notion of a real physical agent, such as a human being capable of performing actions. However, the notion of agent is so general that it can include almost any physical being capable of performing actions, even physical beings without minds, such as viruses. Among the properties generally attributed to an agent are the ability to: (1) act in an environment, (2) communicate with other agents, (3) have individual objectives (driving its behavior), (4) possess resources to perform actions, (5) perceive its environment (though it has only a partial representation of it), (6) possess skills needed to perform actions, and (7) offer services using those skills to other agents to satisfy certain goals. Simulating these seven properties or characteristics in a software system is the subject of *multi-agent systems theory*.

In the 1970s, software agents were fairly simple constructions, typically consisting of a bundle of code that functioned to look for some condition in the system. If the condition exists, then the agent does something. If not, the agent continues to monitor the local situation for the condition. The tasks they perform can be simple or extraordinarily complex, depending upon the program that they run. For example, daemons (that is, agents with a colorful name) in artificial intelligence programs for, say, reading comprehension, might look for words of a certain kind, such as nouns that picked out mammals. If they detected such a noun, then they would either execute additional code or else send that information to some other component of the software system. More recently, complex software systems, such as networked computing systems that have a vast range of practical applications, from electronic commerce to digital libraries, employ subsystems, or components, that act and function as agents. The study of these complex computing structures is part of multi-agent systems theory.

Multi-agent systems can also provide both the architecture and the theoretical structure for real-world nano systems in which the nano-sized agents are robotic-like structures that are controlled by software instructions. Multi-agent systems came of age in the late 1990s. Indeed, by 2002, an enormous amount of work in that field had been done in various research institutions around the globe. We next examine data provided in a paper presented at the Fourth International Conference on Multi-Agent Systems (ICMAS-2000) to demonstrate our claim that ethicists should pay attention to work in multi-agent systems.[15]

[14] Ibid., p. 85.

[15] A special issue of *Artificial Intelligence* (142, 2002) was devoted to many of the papers that were delivered at ICMAS-2000. In a separate paper, we argue that designers of multi-agent systems should pay attention to work in ethics.

Part II: how ethicists can benefit from studying multi-agent systems

Some interesting research has been conducted in AI with respect to the notion of *commitment* in multi-agent systems. In particular, the joint work of Barbara Grosz, Sarit Kraus, David Sullivan, and Sanmay Das on SPIRE (Shared Plans Intention Reconciliation Experiments) stands out as exemplary.[16] We will show that there is an important conceptual connection between the notion of commitment and that of trust, and then show how the work of Grosz et al. can illuminate certain problems concerning the trust relation. This research, in turn, is invaluable for ethicists concerned with the notion of trust. We next consider why this is the case.

The philosophical importance of SPIRE

First, we should note that there are two different kinds of problems concerning the trust relation. One is "philosophical" in the sense that it is concerned with arriving at the correct conceptual analysis of the notion of trust. The other problem is concerned with how to apply the notion of trust in concrete situations— we refer to this as the "technology of trust." For instance, a question about how we should design a diffuse, default trust environment to maximize the trust relation is a question about how to apply the notion of trust in a concrete situation—i.e., the technology of trust. We should also note another important distinction having to do with "concept possession" that is useful for our purposes, viz., a well-known semantic distinction between "meaning-constituting beliefs" and "auxiliary beliefs." Whereas the former are necessary for genuine concept-possession, the latter are not. That I believe a tiger is an animal is necessary for my possessing the concept of tiger; however, it is not necessary that I believe a tiger is an endangered species in order to possess the concept of tiger.

Although some critics, including Quine, may think this distinction is spurious,[17] we do not. We believe that one important task of philosophy is philosophical analysis, which includes the elucidation of the meaning-constituting beliefs necessary for genuine concept possession. Thus, one philosophical question about the notion of trust is concerned with identifying the meaning-constituting beliefs necessary for possession of the concept of trust. Auxiliary beliefs about the concept of trust, however, are not a purely philosophical matter, but instead affect the "technology of trust."

Next, we turn to the work of Grosz et al., which we believe is important in two different ways. Not only does it provide us with a philosophical analysis of the concept of trust, it also provides information about the technology of trust. The philosophical question about trust concerns the conditions under which trust is broken. The "technological" question about trust concerns how the loss of authority of norms may occur unevenly in a community.

Conceptual connections between commitment and the trust relation in SPIRE

SPIRE is an experimental system that models an AA's intention reconciliation and commitments in the context of collaborative activities. It also shows how commitments and intention reconciliation are affected by social norms and social consciousness, as well as by environmental factors and "agent utility functions." What AA's intend to do is what they decide to do, where their decision-making is constrained by the norms of rational decision-making, as well as by resource-bounded reasoning. The significance of the latter is that opportunities and environmental changes may induce an AA to change its plans (where a plan is simply a sequence of actions or intentions to perform some action). The notions of trust and responsibility are not formally defined in SPIRE, though the words surface occasionally in Grosz et al. (e.g., when they assert … "the intuitive notion that teams of agents would be more likely to entrust their most valuable group-related tasks to collaborators who had been most responsible in their past interactions with the group." [18])

For Grosz et al., the notion of commitment is formally defined. Can their notion of commitment be a surrogate for the notion of trust—in other words: When *A* trusts *B* to do some task, is *B* committed to doing it and is *A* committed to *B* doing that task? Can *B* be responsible, but not committed, to doing it? Similarly, can *A* normatively expect *B* to perform the task, though not be committed to *B* that *B* perform the task? If either is the case, then the notions of commitment and trust come apart in important ways, and the work in SPIRE will not help in providing a philosophical analysis of the concept of trust. In the *American Heritage College Dictionary*, a primary meaning of the transitive verb 'to commit' is "to put in trust or charge; entrust."[19] That is, when *A* entrusts *B* to do some task,

[16] Barbara Grosz, Sarit Kraus, David G. Sullivan, and Sanmay Das, "The influence of social norms and social consciousness on intention reconciliation," *Artificial Intelligence*, 142 (2002), pp. 147–177.

[17] However, Quine does allow that a non-principled defeasible form of the distinction can be entertained. See "Two Dogmas of Empiricism," in W. V. O. Quine, *From a Logical Point of View* (Harvard University Press, 1953), pp. 20–46.

[18] Grosz et al*., op cit*., p. 152.

[19] *American Heritage College Dictionary*, *op. cit.* Some take the notion of "entrusting" to be weaker than trusting; i.e., *A* can entrust *B* to do *X*, without *B's* being committed to doing *X*. However, we will not pursue this distinction here.

B commits to doing it, and is responsible for doing it. Of course, the dictionary fails to tell us anything interesting about the relation between commitment and responsibility. But it is clear that there is no notion of commitment that lacks the notion of trust: to be committed is to be entrusted.

We have already seen that the Strawson-Holton-Walker conception of trust adds to the minimal set of necessary and sufficient conditions found in the dictionary the normative notions of responsibility and normative expectation. We next need to consider whether the notion of commitment as used by Grosz et al. can accommodate these normative components. We already noted that one might argue that being committed to doing X does not imply that one is responsible for doing X. And we should further note that one might also argue that being committed to someone to do X does not imply that one normatively expects him to do X, simply because the notion of trust has been assigned a special moral status that is independent of the conceptual connection of trust to commitment. However, it would be semantically odd to say that Jones is committed to doing such-and-such, though Jones is not responsible for doing it. Similarly, it would be semantically odd to say that Smith is committed to Jones to do such-and-such, but Smith does not normatively expect Jones to do it. We take these two linguistic intuitions to be decisive against the objection that the notions of trust and of commitment come apart in philosophically important ways. In particular, when *B* commits to performing a task *B* is responsible for performing it, and when *A* is committed to *B* to performing the task, *A* normatively expects *B* to perform it. In this way, the notion of commitment preserves the normative features of the Strawson-Holton-Walker conception of trust.

In SPIRE, it is the *social norms* that endow the notion of commitment with normative status. In the absence of those norms, the notion of commitment would not align properly with the Strawson-Holton-Walker conception of trust. That is, without normative grounding, *B* could break a commitment to perform a task, and still be trusted by *A* to perform it. But once a commitment is normatively grounded, a breach of commitment would break the trust relation between *A* and *B*.

One might object that in a diffuse, default trust environment, *A* is not committed to *B* to perform a particular task, since *anyone* might perform that task. But the objection is specious, since *A* might (i) be committed to each person (where *A* knows each person), or (ii) be committed to anyone to perform the task, where 'anyone' could be a generic individual. Commitments to generic individuals are dispositional in the sense that when an actual individual comes along to perform the task, *A* is committed to her to perform it.

Another objection to our claim that SPIRE provides important information for a philosophical analysis of the concept of trust can be expressed in the following way: We cannot read results from the data of the SPIRE experiments that apply to human agents in default, diffuse trust environments, since we will never understand if the reasons that AAs provide for their actions are the same reasons humans provide for their actions, even where the reasons are expressed by the same set of sentences. This is to adopt a strong skeptical view toward our use of SPIRE. We will not attempt to refute this view here. However, SPIRE describes such notions as intention and commitment at a level of generality that abstracts away from features that might distinguish AA's from human agents. For instance, SPIRE does not assume that intentions must be manifested at the level of consciousness. Moreover, elementary inferences made in SPIRE—such as computing the expected utility of an action—would not be importantly different from the same inferences made by a human agent, provided the method of computation is the same for both AAs and human agents.

Intention reconciliation in SPIRE

When a human agent decides to adopt a course-of-action under certain conditions, then she has established an intention to pursue that course-of-action. This kind of connection can be simulated in artificial agents.[20] If an agent, either human or artificial, adopts an intention to do task A at time *t* (or within a time interval of the appropriate width), sees that doing task B might be more beneficial than doing task A, and it is the case that doing task A and doing task B conflict (in the sense that at *t*, or within the time interval of the appropriate width, no agent can feasibly perform both A and B, then the agent cannot perform both A and B. In this case, the agent must decide *between* doing task A and doing task B.[21] Experiments conducted in SPIRE study the relationship between intentions that are in conflict and the mechanisms that can be employed for their reconciliation. Examples of intention conflicts in diffuse, default trust environments are straightforward and common. Consider that a human agent, Jones, cannot hold the door open (in a crowd) for Smith if he sees that Jackson has caught her heel on a crack in the floor and is about to fall down. Jones must decide between holding the door open and holding Jackson in an upright position. Which commitment should Jones (or any agent) break?

[20] We will not, in the present essay, at any rate, discuss the issue of whether the simulation involved is of one of an authentic intention, nor will we say what the concept of a human intention consists in.

[21] There are additional possibilities for the restrictions on not being able to do both A and B. These include social commitment policies or cultural norms or psychological norms, such as not wanting to appear outlandish in one's actions. We will not investigate this matter here.

Of course, one might object that the aim of SPIRE is to model intentions, which could leave open the question about any relevance it would have for a philosophical analysis of trust. Whatever an AA in SPIRE decides to do—where the decision to do something is arrived at via reasoning in a decision-theoretic framework—is what that AA intends to do. Thus defined, intentions in SPIRE are fairly tame, and the issue of whether intentions in SPIRE are genuine simulations of human intentions simply does not arise. The substance of SPIRE is investigating how commitments are broken, and, as we have argued above, the notion of a commitment in SPIRE is relevant to the philosophical analysis of trust.

The basic structure of SPIRE and diffuse, default trust environments

The basic structure of the SPIRE framework includes four key elements: sets of (i) agents that can be collected into teams; (ii) team activities, where a team activity is a set of tasks performed by various agents belonging to the team; (iii) incomes that agents receive for the tasks that they perform; and (iv) times during which the tasks are performed. Agents can receive outside offers, which are tasks that can conflict with team tasks, in the following way: An agent cannot perform both a team task and an outside-offer task. If an agent decides to perform an outside-offer task, he defaults on performing his assigned team task, in which case the team must find a replacement agent to perform the assigned team task. Finding replacement agents incurs team costs, which are divided equally among the agents on the team. Agents compute how much income they will receive from their participation in a team, as well as from outside offers. These calculations are subject to constraints that take the form of social commitment policies, of reputation estimates, and of estimates of their social consciousness (brownie-point estimates[22]).

Is there sufficient similarity between the basic structure of SPIRE and diffuse, default trust environments to warrant our claim that we can investigate philosophical problems concerning the notion of trust by employing the data from SPIRE experiments? Since the basic structure is described at a fairly high level of abstraction, it can be used to model any kind of diffuse, default trust environment.[23] A set of agents collected into a team models, for example, human beings in a movie theater, all of whom have gone there for the purpose of viewing a motion picture. A set of team activities models the various tasks that people in a movie theater perform, such as refraining from speaking loudly during the film, letting someone who wishes to leave their seat get past them (while they are seated), and so on. A set of incomes that agents receive for the tasks they perform can model the social stability that arises and personal satisfaction that occurs when people perform those tasks that are necessary for groups of people to interact in a structured social setting in a harmonious way. A set of times during which the tasks are performed models the time during which people are together in a movie theater to watch a motion picture.

A diffuse, default trust environment (such as the Newark International Airport, for example) is typically a dynamic set of distributed relationships. Human agents come and go in a changing physical environment. In such dynamic environments, when a human being breaks a commitment to perform some task, there are usually other people who can serve as replacements, provided that the set of tasks each person performs is not that large in size. In such dynamic environments, though agents rarely know one another, they nonetheless have commitments toward one another in the form of normative expectations and responsibilities with respect to relatively mundane matters, such as walking on the correct side of the street or holding a door open for the person immediately behind them in a crowd exiting a theater. A commitment can be broken in a diffuse, default trust environment—in the language of SPIRE, an agent defaults on a commitment—without breaking the trust relation. If someone fails to hold a door open for Jack, it would be (*ceteris paribus*) irrational for Jack to claim the trust relation manifested in that environment had been broken.

We can envisage several different kinds of contexts for commitments. For instance, there are personal commitments in trust relations of various kinds: love relationships, friendship, family relationships, neighbor relationships, community relationships. There are also work commitments, social commitments, and cultural commitments. In SPIRE, social commitments are modeled in ways that reveal interesting facts about the nature of social commitment and its relationships and interactions with other features, some of which exist outside the context of the commitment.[24]

[22] Grosz et al., *op cit*., pp. 156–157.

[23] Moreover, this level of abstraction isolates constitutive and logical properties of the notions, and thus does not succumb to the fallacy of identifying two different things because they share certain features.

[24] The SPIRE framework includes various parameters that need to be set. These parameters are intrinsic to the kinds of decisions that software agents in a collaborative setting make. In varying the parameters, one can see what the overall effects are on the type of issue studied with respect to that particular parameter. Of course, the ways in which parameters interact in a complex system can be daunting. Thus, the methodology is to vary one or two parameters at a time, to keep the study of the system reasonably tractable. There is also a fundamental distinction between parameters that are intrinsic to the agent and the tasks that agent performs, and parameters that are intrinsic to the environment and the physical setting in which the

SPIRE models two different kinds of social commitments and two different contexts. The two kinds of social commitments are the ranking-based social-commitment policy (RSCP) and the discount-based social-commitment policy (DSCP). The two contexts are sets of homogeneous agents and sets of heterogeneous agents. No matter what the context, there is some point at which the trust relation is broken when sufficiently many instances of a particular kind of commitment have been broken. For instance, when *B* is responsible for performing a task in the context of a trust relation, and has failed to perform the task n times, the trust relation is broken provided that in failing to perform the task n times, *B* abrogates her responsibility to perform the task. Two questions arise: (i) what is the interval in which *n* lies? and (ii) what is the relation between failing to perform a task n times and the abrogation of responsibility? With respect to (i), it is unlikely that there is a single number that signals when the trust relation is broken with respect to an abrogation of responsibility, though we will not argue for that view here.

We contend that the conditions under which trust is broken are an essential property of the concept of trust, since we need to know not only the conditions under which it occurs, but also the conditions under which it is broken, in order to say when it is legitimately instanced and when it is not. But in order to say when trust is broken, we need to characterize when both normative expectations and responsibility in the context of a trust relation break down. It is not enough to say only that the trust relation breaks down when an agent defaults n times, since one might normatively expect an agent to perform some task, without expecting the agent to perform that task. However, at some point not only will an agent not expect some other agent to perform a task—she will also not normatively expect that agent to perform it. What is the character of the relationship between not expecting an agent to perform a task and not normatively expecting an agent to perform a task? Elucidating this characteristic is to provide one key component in a philosophical analysis of trust.

RSCP and DSCP: the social commitment norms in SPIRE

We have already noted that Grosz et al. distinguish two different kinds of social-commitment policies. One kind of policy—the ranking-based social-commitment policy (RSCP)—uses the past behavior of agents to rank them in terms of how often they have defaulted on commitments. (Here, defaulting can occur in different ways—such as taking on another commitment or failing to do the commitment task properly. Agents can also be measured or ranked in terms of how often they have defaulted and on the kinds of offers for which they have defaulted). RSCP norms, in effect, control the opportunities for income available for agents. Where an agent, X, defaults sufficiently many times, there are fewer tasks that X can perform for income than if X had not defaulted. Another way of making the point: an agent, X, loses potential utility increases where X defaults too often.

The other kind of social commitment policy—discount-based social-commitment policy (DSCP)—controls the benefits an agent gets from breaking commitments. DSCP norms are based on three factors: (a) the reputation of that agent as someone who commits in the past, (b) the stability of the group as a whole (or the reputation of the group as a whole), and (c) a social consciousness factor that measures the utility of being a "good guy." While RSCP norms control potential utility-generating factors, DSCP norms control actual utility increasing factors. That is, agents who default sufficiently many times will lose N% of their income from performing team tasks. Notice that there is a spectrum of parameter values for utility-increasing factors, whose endpoints are actual (and no potential factors) on the left and potential (and no actual) factors on the right. Similarly, there is a spectrum of parameter values for reasoning about whether to break commitments, whose endpoints are purely intra-agentive considerations on the left and purely inter-agentive considerations on the right.

Both RSCP and DSCP are relevant to issues about normative expectations and responsibilities in diffuse, default trust environments. For RSCP norms, agents that have been irresponsible in the past over many occasions (that is, agents who have defaulted on their commitments sufficiently many times) will not be trusted in the future as much as agents who have been trustworthy in the past. The opportunities for income available to agents that are controlled by RSCP norms are a measure of the strength of the normative expectations toward that agent. The fewer opportunities that agents have, the less responsible they are as agents (since the opportunities they have are a measure of their past defaulting behavior).

An AA in SPIRE will default on a commitment when it receives an outside offer to perform a task that is incompatible with performing one (or more) team tasks to which the agent is committed and it decides (using a model of rational decision-making) that there is more utility in defaulting than not. The question, then, is what corresponds to an outside offer in a diffuse, default trust environment?

Footnote 24 continued
agent is situated. It is the separation of these two distinct kinds of parameters that allows one to study how each can produce effects within the system and allow one to recognize the cause of those effects. One can, thus, separate out the contributions of a social consciousness factor, such as an agent's reputation, from extrinsic factors, such as the number of job offers that are made to an agent who has a prior commitment to working cooperatively to achieve a particular goal.

Recall that the analogue of income in a diffuse, default trust environment (such as a movie theater) is the social stability that arises and personal satisfaction that occurs when people perform those tasks that are necessary for groups of people to interact in a structured social setting in a harmonious way. An outside offer in this context, then, is the extra utility from breaking a commitment, such as not using a cell-phone during the screening of a film in a movie theatre. Suppose you have reason to believe that a job offer will be made to you by phone during a certain time-period which overlaps with the screening of the film. When your cell-phone rings in the theater, you default on your commitment not to use a cell-phone while a film is being shown. An outside offer, then, is any reason any agent might have to default on a commitment. Conversely, an RSCP or a DSCP norm is a reason an agent has not to default on a commitment. Thus, these social norms provide reasons for agents (whether human or AAs) to have certain normative expectations or responsibilities.

For DSCP norms, the amount of income that an agent receives for the team tasks that it performs is a measure of how responsible it has been in the past. That is, the more often an agent defaults in the past, the less income it receives, and so the less responsible it is inferred to be by other AAs who are members of the team. Moreover, the amount of income the agent receives also measures the extent to which the other agents normatively expect that agent to perform her team tasks. The more often that agent has defaulted in the past, the weaker the normative expectations the other agents have that the agent will perform her tasks in the future. It should be noted that AAs in SPIRE have the capability to infer that an agent has defaulted using the data structures available in the basic structure of SPIRE. In a diffuse, default trust environment (such as a movie theater), a DSCP norm would provide social sanctions for cases in which human agents defaulted on their commitments. For instance, someone who used his or her cell-phone in a movie theatre while a film was being shown might risk the social opprobrium of being ushered out of the theater to a chorus of boos and cat-calls.

SPIRE experiments and their significance for the philosophical analysis of trust

The philosophical questions about the concept of trust that we claim experiments in SPIRE address are mainly concerned with the conditions under which trust is broken. In particular, they question how often an AA must default before trust is broken and what the relation is between defaulting and social consciousness where trust is broken. What we can say about those cases in which the commitments in a trust relation are broken is an important and unsettled part of moral theory—the larger issue within which "moral repair" consists. In one set of experiments involving RSCP norms, and in another set of experiments involving DSCP norms, one variable that was controlled had to do with the homogeneity of the agents on a team. Homogeneity of a team in SPIRE is one in which all AAs on all teams have the same values for the parameters that characterize social consciousness and the weight that is given to future income. Thus, each AA "sees" future income in the same way, and each "sees" social consciousness in the same way, as every other AA. Homogenous sets of teams are analogous to human groups of the same ethnicity, social status, economic status, etc.

In one experiment, a team does better performing the most rewarding tasks when the level of social consciousness is greater than a certain threshold and when the weight that agents give to future commitments is low. In this experiment, there are only two environmental factors that are used: the task density (the number of tasks scheduled for each segment of time), and the rate at which outside offers are made. These environmental factors are analogous to changing environmental conditions in which commitments might change, as well as how many commitments an agent can reasonably undertake in a segment of time. However, as the number of outside offers increased beyond a certain cut-off point, the impact of either RSCP norms or DSCP norms diminished, since AAs were able to offset penalties and lower team incomes with the income from the outside offers. That this happens is certainly not unexpected. The virtue of the SPIRE data is that it provides precisely defined conditions under which it occurs. It also raises the following question: Would a higher parameter setting for social consciousness result in fewer defaults as the number of outside offers increases?

Grosz et al. did not conduct an experiment to investigate this key question. Indeed, a shortcoming of their work—at least with respect to our question concerning the philosophical analysis of trust—has to do with the parameter settings for the social-consciousness factor when obtained in experiments involving a fixed outside-offer rate. Thus, a relationship between defaulting, social consciousness, and the RSCP and DSCP norms needs to be explored more thoroughly before we can fill in the nature of the relationship between defaulting and social-consciousness norms.

The question concerning the technology of trust—that the loss of authority of norms may occur unevenly in a community—is also addressed by the previous experiments. Where AAs are homogeneous, the loss of authority of either RSCP or DSCP norms occurs evenly in teams of AAs. This is almost predictable, given that each AA "sees" social consciousness and income (from team tasks and from outside offers) in the same way. But humans in diffuse, default trust environments are not homogeneous in the way that the AAs in these experiments are.

We should also note that Grosz et al. also conducted experiments involving heterogeneous teams of AAs. In these teams, different AAs have different parameter settings both for social consciousness and the weight given to future income. One experiment showed that in heterogeneous teams constrained by RSCP norms, free-riding diminished where there were fewer outside-offer rates. The analog of the free rider in trust environments is the agent who entrusts others to perform tasks for him, but never performs tasks for another agent. Most of us have witnessed situations in which individuals who never hold doors open nevertheless avail themselves of door-openings performed for them. The results of the experiments on free-riding are counterintuitive—e.g., why should it be the case that where there are fewer opportunities for changing one's commitments, free-riding diminishes? The reason is that the model in SPIRE includes a social-consciousness weight that effectively reduces the benefits that free-riders might accrue. This is certainly a plausible explanation of how free-riding is diminished, but its connection with fewer outside offers to change commitments is unexpected and important.

Grosz, et al. conclude that it would not pay to design a social institution with its attendant social norms and social-commitment policies that encourages agents to deviate from a normatively defined standard of responsibility to the community situated within that institution. Perhaps the same is true of diffuse, default trust environments—i.e., that we should develop social norms that discourage free-riding.

One experiment involving RSCP norms shows that different weights given to the social-consciousness factor can result in different sets of benefits to the agents. In particular, agents that accord weight to social consciousness below a certain threshold tend to default at suboptimal rates. This is analogous to those who don't accord much weight to being responsible in diffuse, default trust environments—such individuals typically default from commitments to tasks and thus do not benefit from this high default rate. Under different task-densities there are important qualitative differences in default rates that are intuitively surprising. For instance, the greater the task density, the less often agents will default. But why should that be the case? Clearly, if there are replacements for doing a certain task that an agent should perform, then the agent might default on that task.

However, Grosz et al., found that the availability of replacement agents had no effect on the default rate in high density task environments. Rather, it is, once again, the social-consciousness factor that accounts for the lower than expected default rate. Indeed, this factor accounts for low default rates across task-densities ranging from low to high. These results are important for developing a fuller account of a diffuse, default trust environment, for the design of such environments should pay attention to the kinds of views people have about responsibility to others and about the different ways in which those who shirk their responsibilities are punished. What the experiment shows, however, is that even where there are others who would gladly do what the free rider wishes not to do, the free rider will see that it is not in his best interests to let that happen. That is, it is in the best interests of the free rider to commit to performing those tasks.[25]

Another set of experiments used heterogeneous groups constrained by DSCP norms in which there were different social-consciousness factors, but in these cases, the two environmental factors—outside offer rates and task densities—were varied. Here the results were expected—the lower the parameter setting for social consciousness below a certain threshold, the worse the AAs fared; conversely, where the interval was above the threshold, but not above that interval, the AAs maximized their utility. One important question that remains, however, has to do with what the specific parameter setting (in this case, an interval of values) means in the context of human agents in diffuse, default trust environments. For example, what does it mean to "see" social consciousness in a certain way? The task that remains is to show how the entire range of social-consciousness parameter settings comports with different attitudes that human agents take toward social-consciousness norms, and the relations of those attitudes to utility-increasing factors such as income and outside offers.

We believe that the work of Grosz et al. is rich and important for ethicists who wish to provide both a philosophical analysis of the concept of trust and a fuller account of diffuse, default trust environments. However, as we have seen, many additional experiments need to be performed to answer key questions concerning both the philosophical analysis of trust and the technology of trust. We suspect that Grosz et al. will have to complicate their SPIRE models in some important ways in order to run experiments that would provide further information that is useful to ethicists who are working in the area of moral trust.

Conclusion

In this essay, we defended the "diffuse, default" account of trust advanced by Margaret Urban Walker. We argued that this model of trust can be applied to experiments concerned with understanding the role of commitment in multi-agent

[25] Grosz, et al., *op. cit.*, pp. 157–159.

systems. We tried to show, via our discussion of SPIRE, how ethicists who are concerned with the theory of trust can benefit from examining experiments that model trust in environments and contexts involving artificial agents. However, we also recognize that there are many open problems involving the concept of trust that were unable to be examined in the present essay. In a separate paper on trust, we specifically address some of these problems.

Acknowledgments We are grateful to Frances Grodzinsky, as well as to two anonymous reviewers, for some helpful suggestions on an earlier draft of this essay.

References

American Heritage College Dictionary. 4th edn. (2002). New York: Houghton Mifflin Company.

Baier, A. (1986). Trust and antitrust. *Ethics, 96*(2), 231–260.

Camp, L. J. (2000). *Trust and risk in internet commerce*. Cambridge, MA: MIT Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Fried, C. (1990). Privacy: A rational context. In M. D. Ermann, M. B. Williams, & C. Guitierrez (Eds.), *Computers, ethics, and society* (pp. 51–63). New York: Oxford University Press.

Gambetta, D. (1998). Can we trust trust? In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 213–238). New York: Blackwell.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2009). Developing artificial agents worthy of trust: Would you buy a used car from this artificial agent? In M. Bottis (Ed.), *Proceedings of the eighth international conference on computer ethics—philosophical enquiry (CEPE 2009)* (pp. 288–302). Athens, Greece: Nomiki Bibliothiki.

Grosz, B., Kraus, S., Sullivan, D. G., & Das, S. (2002). The influence of social norms and social consciousness on intention reconciliation. *Artificial Intelligence, 142*, 147–177.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72*, 63–76.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.

Lim, H. C., Stocker, R., & Larkin, H. (2008). Review of trust and machine ethics research: Towards a bio-inspired computational model of ethical trust (CMET). In *Proceedings of the 3rd international conference on bio-inspired models of network, information, and computing systems*. Hyogo, Japan, November. 25–27, Article No. 8.

Luhmann, N. (1979). *Trust and power*. Chichester, UK: John Wiley and Sons.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21.

Murdoch, I. (1973). *The black prince*. Chatto and Windus.

Nissenbaum, H. (2001). Securing trust online: Wisdom or oxymoron. *Boston University Law Review, 81*(3), 635–664.

O'Neill, O. (2002). *Autonomy and trust in bioethics*. Cambridge, MA: Cambridge University Press.

Quine, W. V. O. (1953). Two dogmas of empiricism. In W. V. O. Quine (Ed.), *From a logical point of view* (pp. 20–46). Cambridge, MA: Harvard University Press.

Simon, J. (2009). MyChoice & traffic lights of trustworthiness: Where epistemology meets ethics in developing tools for empowerment and reflexivity. In M. Bottis, (Ed.), *Proceedings of the eighth international conference on computer ethics—philosophical enquiry (CEPE 2009)* (pp. 655–670). Athens, Greece: Nomiki Bibliothiki.

Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood and agency. *Ethics and Information Technology, 8*(4), 205–213.

Strawson, P. F. (1974). Freedom and resentment. In P. F. Strawson (Ed.), *Freedom and resentment and other essays* (pp. 1–28). New York: Routledge.

Subrahamanian, V. S., Bonatti, J., Dix, J., Editor, T., Kraus, S., Ozcan, F., et al. (2000). *Heterogeneous agent systems: Theory and implementation*. Cambridge, MA: MIT Press.

Taddeo, M. (2008). Modeling trust in artificial agents, a first step toward the analysis of e-trust. In *Proceedings of the sixth European conference of computing and philosophy*, University for Science and Technology, Montpelier, France. Reprinted in C. Ess, & M. Thorseth (Eds.). *Trust and virtual worlds: Contemporary perspectives*. Bern: Peter Lang (in press).

Taddeo, M. (2009). Defining trust and e-trust: from old theories to new problems. *International Journal of Technology and Human Interaction, 5*(2), 23–25.

Walker, M. U. (2006). *Moral repair: Reconstructing moral relations after wrongdoing*. Cambridge, MA: Cambridge University Press.

Weckert, J. (2005). Trust in cyberspace. In R. Cavalier (Ed.), *The impact of the internet on our moral lives* (pp. 95–120). Albany, NY: State University of New York Press.