2020

A Revision of the Buechner–Tavani Model of Digital Trust and a Philosophical Problem It Raises for Social Robotics

Jeff Buechner
CUNY Graduate Center

information

MDPI

Article

A Revision of the Buechner–Tavani Model of Digital Trust and a Philosophical Problem It Raises for Social Robotics

Jeff Buechner [1,2]

[1] Department of Philosophy, Rutgers University-Newark, Newark, NJ 07103, USA; buechner@newark.rutgers.edu

[2] The Saul Kripke Center, CUNY, The Graduate Center, New York, NY 10016, USA

Received: 16 December 2019; Accepted: 13 January 2020; Published: 16 January 2020

Abstract: In this paper the Buechner–Tavani model of digital trust is revised—new conditions for self-trust are incorporated into the model. These new conditions raise several philosophical problems concerning the idea of a substantial self for social robotics, which are closely examined. I conclude that reductionism about the self is incompatible with, while the idea of a substantial self is compatible with, trust relations between human agents, between human agents and artificial agents, and between artificial agents.

Keywords: artificial agent (AA); trust; self-trust; Buechner–Tavani model of trust; self-identification; reductionism about the self; substantial self; personal identity; normative expectation; diffuse default trust

1. Introduction

The Buechner–Tavani [1–3] model of trust is one of the first models of trust that accommodates both human agents and Artificial Agents (AAs)—otherwise known as digital agents (see Taddeo [4]). That is, within the model the following three trust relations can be defined: (i) trust between human agents, (ii) trust between human agents and AAs, and (iii) trust between AAs. Because the model can define these three different kinds of trust relations, it is a model of digital trust, since digital agents—AAs—can have trust relations. However, the model is incomplete in one way—it does not say anything about self-trust. In this paper, the Buechner–Tavani model of trust will be amplified by two additional conditions, each involving self-trust. The addition of these two conditions creates a philosophical problem for social robotics—notably, the problem of a substantial self vs. reductionism about the self. This specific problem is an instance of a more general problem for social robotics: there are philosophical concepts necessary for making certain arguments in social robotics and these concepts are controversial—there are opposing sides as to whether they are or are not viable concepts. In such cases, how do we proceed? Do we simply bracket the point of making an argument to establish some claim in social robotics, or do we attempt to refute one view as to whether the concept in question is or is not viable?

I will examine the more general problem by examining the specific problem, which is an instance of the more general problem. This is the problem of whether, and how, human agents and AAs could have a substantial self (and the attendant concept of a substantial self) in virtue of which they could have (or fail to have) self-trust. I will argue that even though there might be computational realizations of reductive views of the self (such as Parfit's relation R) in AAs, there is an incompatibility between reductionist views of the self and trust. Additionally, even though there might be no computational realization of a substantial self in AAs, a substantial self is needed for trust relations. The main points of this paper are to emend the Buechner–Tavani model of trust and to show that trust is incompatible with a reductive view of the self and compatible with the existence of a substantial self. It is left to another paper to show how AAs can have trust relations with human agents and other AAs even if there is no computational realization of a substantial self in AAs.

2. Adding New Conditions to the Buechner–Tavani Model of Trust

New conditions need to be added to the Buechner–Tavani model of trust [1–3]. The new conditions incorporate the concept of self-trust. The motivation for these new conditions is that human agents must trust themselves in order to trust others. Without trust in oneself, a human agent cannot meaningfully trust other human agents. There are several different reasons for the necessity of self-trust in the Buechner–Tavani model of trust.

2.1. Why Self-Trust Is a Necessary Condition for Trusting Others

2.1.1. Changes in Personal Identity over Time

One reason why self-trust is necessary for trust is that we change over time, often to such an extent that we appear to be a different person at t_n than the person we were at t_{n-k}. We exhibit one form of self-trust when we trust that the person we will later become is one that we will want to become. Similarly, we trust the person we now are to be someone that will take reasonable steps to become the person we will later become. If we were unable to have these relations of self-trust between earlier and later versions of one's self, we would not be able to trust another human being, nor would we understand what it means to trust another human being. Why is that? If we cannot trust our own self to take the steps necessary to become the person we wish to become, how could we trust another human being to take steps to do what we wish them to do? And if we could not do that, how could we understand what it is to trust someone else to take steps to do what we wish them to do?

2.1.2. Self-Trust and Self-Competence

The following argument occurs in a different form in Lehrer [5], where he argues that self-trust is a necessary condition for any human agent to have competence of a certain kind—reasonableness, and joining in the life of reason. Anyone—in Lehrer's view—not worthy of their own trust cannot enter into the life of reason. I am not making that argument here. Rather, I am using the argument to argue for self-trust as a necessary condition for trust (see also Wright [6]).

Suppose that A successfully teaches B how to use modus ponens (MP) and that B accepts what she has learned from A. That is, B conforms to the rule of MP. B then reasons in accord with MP. In this way, B is worthy of the trust of A, for A normatively expects it of B that B will correctly use MP in reaching conclusions in chains of reasoning involving MP. Indeed, it is reasonable that A accepts the conclusions that B draws using MP.

Now suppose that—for whatever reason—B does not understand MP. Although B uses it correctly, B does not have any reason to believe that she is using MP correctly—even though she does use it correctly. B is ignorant of the justification of the rule and thus of its merits. If so, the conclusions that B draws from the rule are not reasonable for B to accept. B is not worthy of her own self-trust because the reasoning that B performs is not reasonable for B to accept, even though it is reasonable for A to accept.

A knows the rule MP and also knows the justification of MP and so the merits of MP. A observes B using the rule MP and knows that B has made no mistakes in many instances of using MP. The reasoning that B makes using MP is reasonable for A to accept. However, it is not reasonable for B to accept, since B does not know the justification of MP nor its merits. B is in that way ignorant of MP.

Now suppose that C explains the rule MP to B—its justification and its merits. However, before B can accept what C tells her, B has to decide whether C is worthy of her trust. If she is worthy of her trust, then B can accept what C tells B about MP. However, if C is not worthy of her trust, then B is not reasonable if she accepts what C tells her about MP. It is only reasonable for B to accept what C tells her about MP if C is worthy of her trust.

To decide that C is worthy of the trust of B, B must be worthy of her own trust. If B is not worthy of her own trust, then any conclusions that she draws and accepts about C are not reasonable for her to accept. It is only if she is worthy of her self-trust that she is reasonable in accepting the conclusions that she draws and accepts about C—in this case, what C has to tell B about the justification and merits of MP.

It might be objected that B is in the horns of a dilemma for which there is no means of escape. The first horn is that B cannot be worthy of her self-trust unless she can justify and know the merits of MP. The second horn is that B cannot accept what C tells her about the justification and merits of MP unless she is worthy of her self-trust. This dilemma also creates a circle: B can be worthy of her self-trust if—and only if—B can be worthy of her self-trust.

Here is how B can escape this dilemma and the circle that it engenders. Throughout her life, B will engage in various episodes of reasoning. Let's suppose that the very first episode of reasoning she engages in is R. There are then two cases: (i) B simply does R and (ii) B learns how to do R from a teacher, T. Let's consider case (i) first.

Is B worthy of her self-trust before she engages in this first episode of reasoning? If no, then it is not reasonable for B to accept this episode of reasoning. If yes, then it is reasonable for B to accept this episode of reasoning. Please note: if B is mistaken about the cogency of this episode of reasoning, she can be corrected by others or even by herself at some future time. That B is worthy of her self-trust does not mean that B cannot be mistaken about, say, this episode of reasoning. Being worthy of one's self-trust does not imply that any of the things about which it is reasonable to accept, one must be veridical.

You might wonder: how did B come to be worthy of her self-trust? This can happen in any number of distinct ways. One way in which it can happen is by episodes of perception of the external world. B sees a tree and accepts that she sees a tree because she believes that she really does see a tree and that there is nothing strange about this episode of seeing. B sees the tree and is not told by someone else that there is a tree in the place where she sees a tree. Her seeing is an intrinsic process—intrinsic to B. In virtue of having this process, B is worthy of her self-trust. She is reasonable in accepting that there is a tree that she is now seeing.

Let's now consider case (ii) above. B learns how to do R from a teacher. B comes to the teacher worthy of B's self-trust, where that worthiness occurs in any number of ways, such as the one described immediately above—the case of veridical perception. The interactions between B and her teacher will begin with B accepting what the teacher tells her because T is a teacher and a teacher teaches their students truths about the world. That is, T is worthy of the trust of B because (a) B is already worthy of self-trust owing to her veridical perceptions of the external world and (b) B has been told by her parents—whom she trusts—that T is a teacher and teachers teach truths about the world. The teacher also knows this, and T knows that both B and her parents know, as do B and her parents know that T knows this. Thus it is common knowledge between T, B and the parents of B that teachers teach truths about the world.

A, B, C and T above are human agents. What happens if A, B, C and T are artificial agents? That is, will the same considerations described above also apply to AAs? This is an important question about the role of trust in digital environments in which AAs perform actions, such as making inferences and making decisions. It is easy to see that the interaction above between a human agent A and a human agent B would carry over to AAs. However, the philosophical question that arises here is the following: what conception of self must an AA have in order to be worthy of its self-trust? Can an AA have a thin conception of self—that is, a conception in which the AA simply calls itself a 'self?' Or must there be a thick conception of self? If so, what could such a conception be when described in computational terms? These questions will be addressed below (see also Carr [7]).

3. The Buechner–Tavani Model of Trust

The five conditions in the Buechner–Tavani [1] model of trust are:

(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

To the above five conditions the following two conditions are added:

(vi) In conditions (i)–(v), substitute 'A' for 'B' throughout

(vii) In satisfying condition (vi), A is said to have self-trust

Some points need to be addressed concerning how it is that A satisfies condition (vi). These points need to be addressed even though there are plausible reasons motivating the addition of conditions (vi) and (vii) to the Buechner–Tavani model of trust. The first is: can A at t_0 have a normative expectation (which may be based on a reason or motive) that A at t_n will do those things that A (now) wants A then to do? There is nothing inconsistent in assuming that A can have that normative expectation. Indeed, the moral of the parable of Ulysses and the Sirens is that an agent can have such a normative expectation, even though it might be very difficult to conform to it.

The preceding consideration shows that A having such a normative expectation at t_0 does not logically violate the Buechner–Tavani model of trust. Can there also be a motivation—within the model—for it? Yes—it is the following. Having a normative expectation toward a future state of myself in which I do such-and-such$_1$ makes it easier for me to understand what it is to have a normative expectation toward another person that they do such-and-such$_2$.

The second point: Is A at t_n responsible for what she does at t_n? Here, too, there is nothing inconsistent in assuming that A at t_0 is responsible for doing at t_n what she normatively expects she will do at t_n. It is part of normatively expecting that she will do such-and-such that she is responsible for doing such-and-such. There is also a motivation within the model for this feature. It is that responsibility for doing what one normatively expects of oneself commits one to doing it. Without being responsible for it, one is hardly committed to doing it—and so it makes little sense to say that one normatively expects it of oneself.

The third point: can A have the disposition to expect that her future self will do those things responsibly? If A at t_0 is responsible for what she does at t_n, then it follows that at t_n she is still responsible for what she at that time does. Of course, there might be various mitigating circumstances under which she is no longer responsible for what she does at t_n, but in the absence of any mitigating circumstances, she is at t_n responsible for what she does at t_n. The motivation for this feature in the model is the same as in the previous consideration.

The fourth point: can A be mistaken about what she will do in her future? Yes—and being mistaken about what she should do in her future is a mitigating circumstances which removes her from being responsible for what she does at t_n. Without the mitigating circumstance of being mistaken about what one should do in the future, one's life would be an ironclad nightmare. This feature is important for the model since it answers to the practical ways in which our actions and plans can be thwarted. Room must be made within the model for such contingencies of life.

The fifth point: can A develop a disposition to trust her future self? There is no reason to believe that she cannot do that—providing that A understands the conditions of trust. That is not to say that she has in mind the Buechner–Tavani conditions of trust. Rather, that she has in mind her own understanding of what it is to trust, and knows that she understands this conception of trust. The motivation for including this feature in the model is that without a disposition to trust oneself in the future, one would have to continually renew the conditions for trust in the model in order to establish that they currently hold. Dispositions to trust eliminate having to do that—having to continually re-establish that the trust relation currently holds.

If all of these points are satisfied, it follows that A trusts her future self—i.e., A has self-trust with respect to future actions (that she do such-and-such) that she performs.

Suppose that none of these conditions are satisfied by A with respect to herself. She does not trust herself. Could she then trust B? No—she cannot, since she fails to satisfy the model of trust when she fails to satisfy conditions (vi) and (vii). To put it another way, self-trust is necessary for other-trust. Can there be other-trust without self-trust? That is, is it conceptually possible for there to be other trust without self-trust? If so, there can be different conceptions of trust—not all of which require self-trust as a necessary condition for trust.

Suppose that A cannot have a normative expectation that her future self will do such-and-such. There are two senses of 'cannot have' here:

(1) A could have a normative expectation toward her future self, but does not (now) have it, for whatever reason or contingency.

(2) A cannot (ever) have a normative expectation toward her future self. It is not because she does not understand what it is to have such an expectation. Rather, it is because of something in her—and what might that be? Start with the case where A does not now have a normative expectation toward her future self. Then how do you generalize to the case where for all t, A does not have a normative expectation toward her future self? This can be done only if there is some condition A satisfies which prevents her from forming normative expectations toward her future self. One candidate for such a condition is that A has the belief that her future self will be a different person than she now is, and that she does not care (now) about that future person (such conditions are much less likely to arise for an agent who believes in a substantial self than for an agent who believes in reductionism about the self).

Default Trust, Zones of Default Trust, and Diffuse, Default Trust

The concepts of default trust, zones of default trust, and diffuse default trust, first defined in Walker [8] were extensively developed by Buechner and Tavani [1–3] to apply to human agent-human agent trust relations, human agent-AA trust relations and AA-AA trust relations. This section provides a brief summary of that material—the interested reader should consult [1] for a much richer account. Walker discusses the attitude of trust one might have toward people one does not know (and has never met) in certain well-defined circumstances, such as using a city subway system. You know not only what to expect from other people also using the subway system, you trust that they will do what you expect they will do. This is common knowledge—you expect and trust me to act in a normal way and I expect and trust you to act in a normal way. For instance, I expect and trust that you will leave me enough space on the subway platform to not endanger me and you expect and trust the same of me. Neither of us knows one another—we are generic individuals to one another within the subway system—but we can each have a normative expectation toward the other that they will do such-and–such. In this case, that each of us will not endanger the other on the subway platform—that each of us will provide enough space for the other on the subway platform. The subway system is a zone of default trust, where knowledge of what is normal, what is not normal, what to normatively expect and whom to trust constitute default trust.

Diffuse default trust occurs when we do not have attitudes of trust toward specific individuals (even generic individuals) but toward institutions—such as an airline corporation. Suppose that airline service provided by this corporation is dismal. One might have an attitude of resentment toward that corporation without having an attitude of resentment toward any given individual, nor any generic individual (whether affiliated with the corporation or not). We can not only expect that the airline provides good service, we trust that it will provide good service—so we have a normative expectation toward the corporation that it does provide good service. Rather than say that we normatively expect and trust that the unnamed and unknown individuals who work for the airline will provide good service, we say that we normatively expect and trust that the airline will provide good service. The default trust relation has diffused not just over a large group of people in a well-defined setting, but also over the organization that makes a zone of default trust possible.

Buechner and Tavani [1–3] describe how default trust, zones of default trust, and diffuse default trust can arise in not only human-human interactions, but also in human-AA interactions and AI–AI interactions, such as using AAs in making financial transactions over the Internet. In such default zones (making financial transactions over the Internet), AAs can have trust relations with other AAs and with human agents. AAs can be responsible for doing what human agents and what other AAs normatively expect them to do, and they can have normative attitudes toward human agents and other AAs.

There are multiple, different and independent descriptions of the behavior of AAs in default zones. Instead of the terminology of default zones, default trust and diffuse default zones, one might instead adopt the terminology of computable social and natural constraints and how AAs can compute maximal or even relative maximal solutions to such constraints. It would be an interesting project to compare and contrast the different descriptions, but that is not the goal of this paper. I assume that default trust, zones of default trust and diffuse default trust can be used to reliably describe the interactions of human agents with AAs (such as trust relations between them).

Whether AAs do or do not have a computationally realized substantial self, human agents can interact with them as though they do, provided that AAs are competent to reason about the trust relation and have the concept of a substantial self. AAs in that case can make meaningful statements about a substantial self and about trust. In a default trust environment—such as making a financial transaction on the Internet—in which AAs are necessary components, human agents can engage in trust relations with those AAs. It might be objected: we cannot trust AAs in such a default trust environment—we can only expect them to do such-and-such, but not normatively expect them to do such-and-such. This is no different from expecting that dough will rise in the oven. We do not normatively expect that dough will rise in the oven—it is incoherent to have reactive attitudes toward rising dough.

In response to this objection, it should be noted that in a default trust environment a human agent can normatively expect other "generic" human agents (other human agents not known to the human agent) and AAs to do certain things for them. How this happens is described in much greater detail in [1–3]. That a human agent could have a normative expectation toward an AA not having a computationally realized substantial self (but having the concept of a substantial self in its lexicon) would be a matter of both being in a default trust zone in which trust diffuses to AAs (I leave the detailed description of how this happens to another paper. To provide that description in this paper would significantly increase its length and distract the reader from its purposes).

4. AAs, Self-Trust and Explicit Knowledge of the Buechner–Tavani Model of Trust

Human agents can exhibit a trust relation toward themselves and others even though they do not know the Buechner–Tavani conditions for trust. Indeed, even though they do not know the conditions for any model of trust in the literature. It would be absurd to say that one could not trust oneself nor trust another unless they knew the Buechner–Tavani model of trust—or any other model of trust. Why is that? It is important to answer this question, because we will soon see that for AAs we want them to know the conditions expressed in some model of trust before we can say of them that they can trust both human agents and AAs (including itself). So in this respect AAs are different from human agents with respect to the trust relation.

For human agents, we waive the requirement that they have explicit knowledge of some definition of trust before they can be said to exhibit the trust relation, since human agents can exhibit, say, normative expectations that another human agent will do such-and-such without knowing that exhibiting normative expectations of that kind is a condition in some model of trust—say, the Buechner–Tavani model of trust. If a human agent fails to exhibit a normative expectation toward another human agent that they will do such-and-such, we can say that they are not exhibiting the trust relation toward that other human agent.

Why should we want an AA to have explicit knowledge of some model of trust, even though a human agent will not know it (but will act as if those conditions are satisfied when they exhibit a genuine trust relation)? If the AA does not have explicit knowledge of some model of trust, we—the human agents—cannot trust that AA to reason properly with respect to trust. Without explicit knowledge of some model of trust, a human agent will have no idea what it is an AA is doing, even if their behavior appears to conform to some model of trust. In which case, a human agent could not reciprocate trust toward the AA. On the other hand, we generally can trust another human agent when their behavior is acceptable with respect to a situation of trust.

There are also circumstances when human agents must have explicit knowledge of some model of trust. For instance, when human agents reason about trust, they, too, must have explicit knowledge of some model of trust. We cannot trust that their reasoning is correct unless we also know what model of trust it is that they are using in their reasoning. The model they use might not be anything more than their own intuitive view as to what trust consists in, and that will be sufficient. If one doubts the cogency of their reasoning, one factor responsible for the lack of cogency might be the model of trust they employ in their reasoning.

5. The Concept of a Substantial Self and the Role It Plays in Trust

Since self-trust is a necessary condition for trust, the idea of a substantial self plays an important role in trust relations (see Rosenthal [9] for a discussion of what a substantial self might consist in). There are difference views as to what a substantial self consists in. Some philosophers have argued that the idea of a substantial self is a chimera. Derek Parfit has argued that the idea of a substantial self is a chimera [10]. The multiplicity of views on what a concept of a substantial self consists in—or even whether there is a concept of a substantial self that is philosophically coherent—presents a problem for human agents engaged in a trust relation. Depending on which philosophical view of the substantial self that one takes, one's interpretation of the conditions for trust in the Buechner–Tavani model can differ from one human agent to the next. Unless these differences can be explicitly recognized, two human agents with different views of a substantial self might find that they cannot trust each other even though each appears (to himself) to satisfy the conditions for trust.

The issues which this paper examines are (i) whether an AA needs to have an explicit conception of a substantial self in order to engage in trust relations or an explicit conception of a reductionist view of the self, or no conception of the self at all, (ii) what happens to trust relations when different AAs have different conceptions of a substantial self or of a reductionist view of the self, and (iii) what happens to trust relations when AAs have different conceptions of a substantial self or of a reductionist view of the self and engage in trust relations with human agents.

I will argue below that although a reductionist view of the self could be computationally implemented (because relation R could be computationally implemented over descriptive memory protocols), a reductionist view of the self is incompatible with realizing a trust relation: there could be no genuine trust relation compatible with a reductionist view of the self. At most, that an agent could expect another agent to do such-and-such is compatible with reductionism about the self. However, normative expectations that an agent could do such-and-such are not compatible with a reductionist view of the self. On the other hand, a genuine trust relation is compatible with a substantial view of the self. However, whether a substantial view of the self can be computationally implemented is an open question—perhaps as hard as the question of whether there could be a computational implementation of consciousness. There is currently no argument in the literature that it is impossible for there to be a computational implementation of a substantial self.

Before examining the question whether an AA needs to have an explicit conception of a substantial self in order to engage in trust relations with other AAs and with human agents, we need to examine the question: what kind of idea of a substantial self must a human agent have in order to have self-trust? We start here because if there is an answer that there is some specific kind of concept of a substantial self that is necessary for self-trust, we would want to simulate that concept of a substantial self in an AA, if possible. Although there are many concepts of a substantial self (in philosophy and in various world religions), we restrict this paper to examining two prominent concepts of a substantial self—that of Derek Parfit and the philosophical view that there is a substantial self that is the bearer of everything that a human agent does—such as thinking, perceiving, and acting.

But perhaps the condition that trust requires self-trust is mistaken because there is a circle when self-trust is included in the definition of trust. You cannot define self-trust until you have a definition of trust. However, if self-trust is part of the definition of trust, there is a circle. If so, self-trust cannot, without circularity, be a necessary condition for trust.

This objection can be easily answered: there is no circle when self-trust is a necessary condition for trust. All five of the conditions for trust (excluding the two conditions for self-trust) are specified. Then the self-trust condition is added—and it must satisfy the first five conditions on trust. There would be a circle if the conditions for self-trust had to be fulfilled before the other five conditions for trust.

An Open Question about the Computational Realization of a Substantial Self

I will argue below that although a reductionist view of the self could be computationally realized (because relation R could be computationally realized over descriptive memory protocols), a reductionist view of the self is incompatible with establishing a trust relation: there could be no genuine trust relation compatible with a reductionist view of the self. At most, that an agent could expect another agent to do such-and-such is compatible with reductionism about the self. However, normative expectations that an agent could do such-and-such are not compatible with a reductionist view of the self. So it would do no good to computationally realize a reductionist view of the self in AAs that engage in trust relations with other AAs and with human agents.

On the other hand, a genuine trust relation is compatible with a substantial view of the self. However, whether a substantial view of the self can be computationally realized is an open question—perhaps as hard as the question of whether there could be a computational realization of consciousness. There is currently no argument in the literature that it is impossible for there to be a computational realization of a substantial self. However, do AAs need to have a substantial self in order to reason about and to engage in trust relations with human agents?

I conjecture that even if AAs do not have a computationally realized substantial self, but do have the concept of a substantial self in their lexicon, and are competent to use that concept in reasoning about trust relations, that they can have trust relations with human agents and with other AAs. However, this paper will not address this question head on. In another paper I will show how the Buechner–Tavani model of trust can be extended to diffuse default trust situations in which there are human agent—AA trust relations and AA–AA trust relations and in which AAs do not have a computationally realized substantial self. The aim of this paper is more modes: to show that a reductionist view of the self (which might be computationally realized) is incompatible with satisfaction of the conditions for trust in the Buechner–Tavani model of trust and that a substantial self is compatible with trust relations.

6. Human Self-Trust and AA Self-Trust

Are you worthy of your own trust? If you make good chains of reasoning and good evaluations, you are worthy of your own trust. If you arrive at conclusions through haphazard chains of reasoning or make evaluations that are bad but lucky (or bad and unlucky), you are not worthy of self-trust. The question is—who (or what) is the self who (that) is worthy of trust?

6.1. Parfit on the Substantial Self

On a reductionist view of the self, of the kind formulated by Derek Parfit [10], there is no substantial self. There is no concept of personal identity that is connected to the idea of a substantial self—something that persists over time and in the face of both physical and mental changes. It consists in satisfying relation R—psychological continuity and psychological connectedness. Psychological continuity differs in important ways from psychological connectedness—the latter of which is what John Locke [11] took to be the primary feature in personal identity. Suppose that I can remember what I did yesterday, though I cannot remember what I did two days ago. However, suppose that yesterday I can remember what I did the day before. Call a memory I have of what I did yesterday a direct connection. Then even though I am directly connected to what I did yesterday, and yesterday I was directly connected to what I did the day before, I am not now directly connected to what I did two days ago, since I cannot remember what I did two days ago. Psychological connectedness is not a transitive relation.

But there is overlap between the memories that I have today and the memories that I had yesterday. It is this overlap which makes for continuity of memories. The relation of psychological continuity is transitive, since if there is overlap between today and yesterday, and there is overlap between yesterday and two days ago, then there is overlap between today and two days ago. It is the overlap which makes the relation transitive. The contents of the memories are not what is important in overlap—only that there is an overlap. If A overlaps B and B overlaps C, then A overlaps C. Whether this distinction really captures an essential feature of personal identity (that of transitivity), is another question—Parfit certainly takes it to do just that. Although overlap is all or nothing, that is not the case with psychological connectedness—which can have degrees along a spectrum from all to nothing. Because connectedness is more important than continuity in determining whether some person (or some object) survives over time, and because it comes in degrees, there will be cases in which we cannot tell whether a person has or has not survived. If this is the concept of a self an AA has, then there will be cases in which we cannot tell if an AA has or has not survived, say, a software error. Additionally, these will also be cases in which the AA cannot tell if it has or has not survived.

Here is a simple case. Someone at one time, t_0, might bear almost no degree of psychological connectedness to their later self at t_1. If this happens, then that person at t_1 does not bear the appropriate R relation to the person at t_0, even though the persons at the two distinct times share the same body. If there were a deep further fact, in addition to relation R, that there is a self which unifies all of the memory impressions in relation R, then we would say—and the person would no doubt say—that they are the same at t_0 and at t_1, even though great psychological changes have occurred in the temporal interval between t_0 and t_1. However, where there is no self, this would not be said—it would not be a belief that one could reasonably have about the situation. It would be rational for them, and for us, to believe that the person they are at t_1 is not the same person at t_0. That the two temporally separated bodies are not the same person means that the person at t_1 might have an attitude toward the person at t_0 of them being a different person, and, similarly, the person at t_0, anticipating great psychological changes, could rationally believe that he should not have the attitude toward all future states of his body that they are the same person as he is, now, at t_0.

We might find the same kind of situation in AAs, where an AA at t_1 might not bear the appropriate R relation to the AA at t_0. In the case of AAs, we cannot speak of bodies unless the AA is embodied in an artificial body. An important difference between an AA and a human being is that AAs forget only if there is some software error or some hardware malfunction—otherwise, they do not forget. That is not to say AAs could not be programmed to forget. In either case, though, an AA could fail to satisfy the R relation by intense episodes of forgetting caused by either software errors or a software computed function.

Since the trust relation consists, in part, of a normative attitude that one agent takes toward another agent—that those who trust take toward those they trust—the question that is raised by AAs that fail to satisfy the R relation is what normative attitude they can take toward other AAs and toward themselves. Their attitude toward their earlier self might be like their attitude toward other people. Thus someone now might not have a good reason to trust themselves as they will be later, since their later self might be someone wholly other to them as they are now. Their later self might not be a self to be trusted. Given that one cannot predict how the future will turn out, it is prudent not to trust one's future self, since one cannot predict to what extent one's future self will be someone wholly other. This is a defect of Parfit's theory of the self in the context of the Buechner–Tavani model of trust, which requires that an agent—human or artificial—be able to trust themselves. If there is a deep further fact about one's self, then the worry that a future self will be so different from one's current self that it will be another person is ill founded. However, it is not ill founded on Parfit's theory of personal identity.

Where agents believe reductionism about the self is the case, then it is not clear that they (human agent or AA) can exhibit self-trust, since self-trust requires that it be durable—that it extends over time. The idea of self-trust for the moment does not make sense. "He trusted himself for only an instant." "He trusted himself for only five minutes." These are remarks that might appear in literature, as figures of speech, such as synecdoche. A substantial notion of self-trust must involve self-trust over some substantial temporal interval. How large the temporal interval should be for self-trust to be substantial is not a matter I will address here. It may be that it must be larger than one year.

But when a temporal interval is taken to be a necessary part of self-trust, the possibility of radical shifts in psychological connectedness arise, and with that possibility, the attending possibility that one might not trust their future self. Thus, given that such changes in relation R are possible, it might not be prudent to trust a later self, and thus might not be prudent for one to take on a substantial form of self-trust. However, even if one does not exhibit self-trust, how should that affect trusting someone else?

Recall that trust is a reciprocal relation, and that there are normative expectations for both parties, A and B, in the trust relation. If an agent does not trust himself in the way in which we have envisaged it above, the same agent would be prudent not to trust others, since he cannot guarantee that the reactions and normative expectations he must now manifest to engage in a trust relation (because they are a part of the trust relation) will actually continue in the temporal interval during which the trust relation holds. That is, the possibility that in the future one might no longer have the normative attitudes which are necessary for a trust relation between two persons to obtain shows that it would be prudent not to trust others given this possibility about oneself. So failure of self-trust exerts two distinct strains upon a trust relation: a strain on self-trust as a necessary condition for trust, and a strain upon the normative attitudes which are necessary for trust.

Parfit [10] has argued that there are situations in which what appear to be distinct possibilities as to what is the truth are merely different descriptions of the same set of facts which do not reflect differences as to what is true and what is false. For instance, he argues [10] that in cases of teletransportation in which there is no overlapping of lives (in which there are no branch-lines—it is not the case that there are two possibilities) that the successor on Mars is me or that I have died and it is my replica (each of which is physically mutually exclusive of the other). Consider: you enter a teletransportation device, and your body is destroyed. The blueprints for your body are sent to Mars, and there another body is assembled out of raw materials. It appears there are two possibilities as to what happened: (i) you died in the teletransportation machine and a replica of you was created on Mars and (ii) you did not die in the teletransportation machine—your existence continued on Mars. If (i) is true, then (ii) is false, and conversely. Either you died or you did not die, period.

These are not distinct possibilities as to what is the true state-of-affairs in the world, but rather two different descriptions of the same set of facts. One can, according to Parfit, decide that one description is superior to the other for various purposes, but it is not the case that one description is true, while the other is false. Since they do not mark out possibilities as to the true state-of-affairs, they are not truth-evaluable. Describe the facts of teletransportation, and you know everything there is to know about that situation. Thus, there are no further things to know (about which you might be right or wrong, about which there are possibilities, one of which is true and the others of which are false). There might be, for a given context, an optimal description of the facts. Thus, there might be reasons for preferring the description (i) over the description (ii). Indeed, Parfit takes (ii) to be the optimal description—for it is one in which you survive because the relation of psychological continuity and connectedness is preserved in the body that is created on Mars. However, there might be other contexts in which (i) is the optimal description.

Although these cases are not (now) technologically possible for human agents, there are analogues of them for AAs that need to be addressed. An AA can undergo malfunction and then correction in the same way that a human being can be reconstituted from scratch on another planet. The malfunctioning is the analog of destroying the body on earth and the correction is the analog of being reconstituted (from scratch) on another planet. Do we want to say of such AA cases that there are not distinct possibilities as to what is the true state-of-affairs? Are there just two different descriptions of a single set of facts, neither of which is the truth of the matter, because neither is truth-evaluable? If we do take this position, then we are left with a problem. It is this: what we say of

AAs in such situations is a matter of convention—we can choose one description over another, but neither description that we choose is truth-evaluable.

The problem is that AAs communicate with one another by sending messages that contain valuable information. If there are breakdowns and then corrections in a network of AAs, whose conventions will be used to decide the identities of the various AAs? Without making that decision, no AA will know which AA to send messages to, and that would mean that any algorithm that they implement would not succeed in finishing its computation. If it is not a matter of truth as to the identities of AAs that inhabit a network after a breakdown and the subsequent correction, how would the conventions that are employed to identify AAs possibly be justified? This is a serious problem for a conception of the self which is based on reductionism about the self.

Additionally, what happens to normative expectations a human agent might have about what actions an AA is responsible for doing, if what AA it is the human agent has normative expectations toward is a matter of convention? If I don't know what AA I am supposed to have a normative expectation toward unless I make a convention that it is one rather than the other, can I really have a genuine normative expectation that they will do such-and-such? No doubt Parfit would say that we can—that it does not matter which AA we have a normative expectation toward, since there is no truth claim that attends the claim that it is one AA that is the genuine AA (or conversely). However, if this is how Parfit would respond, we can fault it, on the following grounds.

Suppose that a human agent communicatively interacts with a network of AAs, and that one or more AAs in the network undergo a breakdown and then a subsequent correction. If it is a matter of making a convention as to the identities of the various AAs in the network, can a human agent justifiably have normative expectations toward the AAs in that network? It seems that the identities of the agents should be a matter of fact that is truth-evaluable. Can a human agent justifiably entertain a normative attitude toward an AA whose identity that human agent simply decides by a convention? (See Buechner [12,13] for a discussion of why the decision must be made by a convention). The normative expectation the human agent entertains will be that various AAs do such-and–such and that the AAs are responsible for doing such-and-such. However, if their identities are a matter of convention (because of the breakdown and subsequent correction), it closes off justifying the normative expectations of the human agent because justification is a notion that requires sensitivity to facts, and in this case there is (according to the views of Parfit) no fact-of-the-matter as to the identities of the various AAs in the network.

Another problem for reductionism about the self that arises for human agents is the following: does it even make sense to engage in a participant reaction (such as normative expectation) with nothing more than a connected stream of memories and a body (see Unger [14])? Or is it necessary for there to be a substantial self at both ends—that one reacts to and which is the locus of the reaction? This can create difficulties of various kinds. For instance, if a human agent—whether she does or not believe there is a substantial self—believes that the other human agent might not believe there is a substantial self, can she then have a normative expectation toward that agent that they will do such-and-such? What would impede having such a normative expectation on her part is the possibility that the other human agent might not trust her own future self because she might think that her own future self will be sufficiently different from her current self as to constitute a different person. In which case, she would not be responsible for what her current self normatively expects of her. The first human agent, knowing that this is possible, would then not be able to have a normative expectation that the other human agent will do such-and-such for her at any time in the future, near or far. Thus, that someone might believe in reductionism about the self would create difficulties for forming normative expectations, even where those who would form them do not believe reductionism about the self.

Perhaps there is some surrogate for the normative relations (such as normative expectations), some surrogate of what is a normative relation that allows for an attenuated trust relation, or for trust*, or for quasi-trust. However, just as Parfit claims that there is no such thing as the self, no deep further fact behind psychological continuity and connectedness, so, too, if he is right, and it is the case that the normative attitudes are necessary conditions for trust, there is no notion of trust that will result in stable trust relations between human agents, between human agents and AAs, and between AAs.

6.2.A Possible Counterexample to the Buechner–Tavani Model of Trust

It might be objected that a human agent at t_0 might take their future self at t_n to be so unworthy of self-trust to do certain things that they would trust other human agents or AAs to do those things. In which case, there would be a case of trust that violated conditions vi and vii in the Buechner–Tavani model of trust.

This is an objection to a reductionist view of the self. In a reductionist view of the self, relation R at t_n could be satisfied by some person other than the person at t_0, even though they both occupy the same body at each time. If a reductionist view of the self is incompatible with trust (on grounds other than that conditions (vi) and (vii) of the Buechner–Tavani model are violated), then it is a view of trust that cannot be accommodated within the Buechner–Tavani model. So, an example in which there is trust, but a violation of conditions (vi) and (vii) in the Buechner–Tavani model of trust is not a counterexample to the model, since it is an example in which a reductionist view of the self is presupposed, and that reductionist view of the self is incompatible with the model. On the other hand, a substantial self is compatible with the model—as will be argued below.

The objector might continue, though: "Isn't it possible, even where a human agent has a substantial self, for that agent to consider their future self to be unworthy of self-trust? If so, it would be rational for them to trust other human agents or AAs to do certain things they would not trust themselves to do." In response to this objection, two points need to be made. The first is that where there is a substantial self, it will still be the same person at t_0 and at t_n. It would not be a different person at t_n. The second is that self-trust occurs in respect of being a rational agent and a competent reasoner. If my competence as a reasoner is so bad that I cannot trust myself to reason properly, it would not be rational for me to trust someone else, for my reasoning competence would also make it difficult, if not impossible, for me to understand what it is to trust someone else to do certain things for me. Being worthy of one's own trust is being able to engage in certain basic forms of reasoning, without which one is not capable of reasoning either about oneself or about others. So if there is an example in which conditions vi and vii of the Buechner–Tavani model are violated, it does not follow there is still trust between the human agent in question and another human agent. There will not be trust between those human agents when those conditions are violated, so such examples are not counterexamples to the model.

It is perhaps easier for an AA to be worthy of its own trust than it is for a human agent, since one would not expect the reasoning of an AA to be so unreliable that it is rarely an instance of proper reasoning. However, it can happen that software bugs could make the reasoning of an AA unreliable. However, those bugs can be patched. In this way, the deterioration and subsequent restoration of reasoning would be a problem for reductionism about the self, but not for a substantial self (as I noted above). It would be a problem for reductionism about the self, because the deterioration and restoration might define—using relation R—different AAs. On the other hand, if there is a substantial self—even if it is not computationally realized in an AA, but the AA has in its lexicon the concept of a substantial self—it will be the same AA over time

7. Computational Realization of the Self

An issue in artificial intelligence is: if there is a substantial self, can it be computationally realized? Suppose that reductionism about the self is true—that we do not have a substantial self. Then there is no substantial self to computationally realize—and thus, a hard problem in artificial intelligence disappears. On the other hand, relation R would probably not be difficult to computationally realize in an AA. Someone might take this to be a reason to believe in reductionism about the self. However, even if reductionism about the self is true, there are problems, orthogonal to the ones addressed above. Here is one such problem. Suppose all human agents falsely believe there is a substantial self, and act on the basis of that belief, and interact with AAs. Suppose also that it is computationally possible to simulate a faux substantial self in AAs. If so, then it would be important to computationally simulate a faux substantial self in all AAs that interact with human agents. Then AAs would have faux substantial selves (even though they are false entities) and thus would be capable of engaging in a form of trust that requires the notion of a faux substantial self. Even though both human agents and AAs had false beliefs about the self, it appears that nothing would jeopardize the system of normative relations that make up the trust relations between human agents, between human agents and AAs and between AAs. Stable trust relations would be formed even if the idea of a substantial self is a false view.

But if some human agent came to truly believe in reductionism about the self, and human agents who falsely believed in the self knew about this, the latter would not be able to develop stable trust relations with the former, since they would not be able to have normative expectations that the other agent will do such-and-such. In this situation, those who had true beliefs about the self would not be able to have trust relations with those who had false beliefs about the self. However, among those who held true beliefs about the self, they would not be able to have trust relations among themselves as well. Only human agents who falsely believed they had a self would be able to engage in such forms of a trust relation with AAs (of course, this is what might happen if reductionism about the self is true).

8. Desires and the Existence of a Substantial Self

Can a human desire be recognized in a human agent by another human agent if there is no substantial self that could be the bearer of that desire (see [15])? Certainly, anyone who believes in reductionism about the self could express in words that they have a desire and would certainly "feel" the desire. That is, they would have first-person experience of the desire. However, if there is no substantial self that has the desire, is the desire then 'free-floating'? Is the desire not attached to anyone? If persons are chains of memories related in the right way (by relation R), then it is not clear that we could recognize that someone has a desire other than by inspecting their memories in order to determine which person it is that has that desire. Just how would we do that? All we have to go on are the words that the agent uses to express their memories.

This is a problem for those who believe in reductionism about the self. Those who believe in a substantial self will not be vulnerable to that problem. They can decide what the desires are that those who believe in reductionism about the self say they have by taking those words at face value. There is no need to inspect memories to decide which person it is that has that desire. Suppose that AAs have to establish that a human agent has a specific desire. If the AAs believe in reductionism about the self, then they will not be able to establish that a human agent has a specific desire, for the same reasons that a human agent would not be able to do so.

A simple example would be one in which a human agent A wants to manipulate B into doing something by deceiving B into believing that A trusts B to do such-and-such for A. However, A has no desire to have the action performed, but does have a desire that B perform it because A wants B to perform it. A reductionist about the self would have to determine whether A's memories and B's memories satisfy relation R. This might be a computationally intractable task. On the other hand, those who believe in a substantial self would not have a problem in identifying A as the owner of the desire and of identifying B as the person manipulated by A.

For another example, take a financial transaction on the Internet which is mediated by AAs. Suppose that a human agent A has the desire to make such a transaction, and trusts the AAs in the default zone in which the transaction is made. How would an AA establish that it is human agent A who has that desire if the AA is programmed to reason about human agents and AAs as having no substantial self, but whose memories do or do not satisfy relation R? On the other hand, the AA would have no difficulty in identifying A as the owner of the desire on the view that A has a substantial self (both of these examples can be used, with appropriate modifications, in the sections below).

Indeed, there are many aspects of human mentality that a human agent who believed in reductionism about the self would not be able to recognize, if recognition depended on inspection of a stream of memories. Recognition of many features of human mentality is more than just seeing that a specific memory content is such-and-such (and not something else). For instance, being able to recognize that someone has made a request may also require the existence of a substantial self—that there is an agent who has made the request, and that the request is not simply an utterance that is recorded in a memory. Similarly, can a physical or psychological need be recognized if there is no easily identifiable agent who has that need? If there is only relation R—psychological continuity and connectedness—then what person is it that has a need? When some human agent says that he has a need, but there is no self that is the subject of that need, whose need is it?

Continuing this line of questioning: if there is no substantial self, then does it make sense to react to a human agent who breaks a trust relation, with anger, or disgust, or revenge? Who is to blame—who is the locus of responsibility? If there is just psychological continuity and connectedness and no deep further fact—a substantial self—then there appears to be no one who can be easily identified who can be blamed—no one who is the locus of responsibility. Punishment, blame and responsibility in a community of human agents for whom reductionism about the self is true might become little more than ideas which have no application to those agents. Madell [16] claims that "an analysis of personal identity in terms of psychological continuity and connectedness is utterly destructive of a whole range of our normal moral attitudes … Shame, remorse, pride, and gratitude all depend on a rejection of this view."

Parfit [10] notes that the view that desert and reductionism about the self are incompatible can be defended, as can the contrary view that desert and psychological continuity are compatible—that we can be punished for past crimes. However, he also says that he has not succeeded in finding an argument which shows that one, but not both, of the views can be successfully defended. That is, that there is an argument that shows that, given the two views, only one of them can be successfully defended. What this means is that it is an open question as to whether desert and relation R—psychological continuity and connectedness—are compatible. If so, that means that whether there is a notion of trust that applies to human agents is an open question, but only on the view that reductionism about the self is true. If reductionism about the self is given up, the open question about the compatibility of desert and relation R becomes an irrelevant concern.

9. Motivation and Trust

Walker [8] poses the questions, "To what in our shared situation and with what motivation, do we expect others to respond when we trust them?". If there is no substantial self, then what happens to personal motivation? What are the motives for an action when there is no substantial self which is the bearer of those motives? If there is just a partially connected stream of memories, and it is not the case that they are unified by being *my* memories, it seems hard to conceive of how there could be personal motivation. If there is no substantial self—if reductionism about the self is true—then there is a severe problem for trust. I am motivated to do such-and-such (e.g., forming a normative expectation toward another human agent), because it will result in such-and-such for me (the actions that the other human agent will responsibly perform). This schema can be easily understood when there is a substantial self.

But what if all there is happens to be the satisfier of relation R? In that case, we have 'the satisfier of relation R is motivated to do such-and-such, because it will result in such-and-such for the satisfier of relation R.' This latter schema cannot be easily understood if there is no substantial self that is the satisfier of relation R. Indeed, it cannot be understood, since the idea of motivation is tied to the idea of a substantial self. Without a substantial self, how can we speak of the motives of the satisfier of relation R, which is just a set of memories! How can a set of memories be motivated to do anything at all? For without motivation and a substantial self which serves as the bearer of the motivation, we cannot have a normative expectation that another will do such-and-such, since the various kinds of motivation which we would think to be appropriate to doing such-and-such for us in that context will be absent. Where we think favorably or unfavorably of someone's motives, then we take a reactive attitude toward them and toward their motives. However, where there are no motives, there are no reactive attitudes we can take toward them. Eliminating motivations for actions from the context of a trust relation is to eviscerate trust, since a motivation for performing an action is a central aspect of trusting someone. If I look upon your motivation for doing such-and-such for me unfavorably, I will be disinclined to trust you to do it. On the other hand, I will be inclined to trust you to do it where I look upon your motivation for doing such-and-such favorably.

There is a significant difference between predicting what someone will do where we know what their motive is in doing it, and predicting what they will do when we do not know what their motive is in doing it. The difference consists in the knowledge that we have regarding why they do what they do. Where we do not know the motive, we have less knowledge on the basis of which we can predict what they will do than where we know the motive. However, if a motive is simply viewed as an item of knowledge, then there is no normative aspect to it. Thus, there is nothing to distinguish in kind between prediction on the basis of information of quantity A and prediction on the basis of information of quantity B, where Inf A ≠ Inf B. In either case, it is a prediction based on information.

But in the case of trust, there *is* something more when we know the motives of the other than just having additional information to use in making a prediction about what they will do. What the difference is between what more we have and the amount of information that we have consists in the normative aspects of the information—in the normative aspects of the motives that the other has with respect to doing such-and such in the context of a trust relation. It would be a mistake to say that we do not acquire information in learning of (and subsequently knowing) the motive of the other. However, it is not just information that we acquire—here the mistake is to think that it is just information that we acquire. Rather, we acquire something more—and that is a normative claim of the other to make good on what they say they will do. It is not just saying that they will do such-and-such, but saying it in a way that gives it authority and binds it to us.

10. AAs and Self-Trust

AAs can construct quite intricate chains of deductive reasoning. However, can they make (i) good evaluations and (ii) good chains of commonsense reasoning? So even though their deductive reasoning skills make them worthy of their own trust, their commonsense reasoning skills and evaluative skills might not. This is not a problem for human agents, since we do not need to produce a computational account of our skills in order to justify them to ourselves. That we succeed in using them in social contexts is usually the means by which we determine that we have those skills. What shall we say about AAs? What level of skill must they have before we can say that they are worthy of self-trust? Should it be determined in the same way that we determine it for ourselves—namely, in a social context? Or do we require a computational account of those skills before we can say that they are exhibited in AAs? Notice that in order for AAs to exhibit those skills successfully in social contexts, there must be a computational account—otherwise, the AAs would not know what to do in social contexts.

This creates a problem for using the Buechner–Tavani model of trust for AAs. If we do not have a computational account of evaluative skills and commonsense reasoning skills that are used in social contexts, we cannot say whether the AAs have and successfully employ those skills. In which case, an AA is not worthy of its own trust.

One way around this is to limit the chains of reasoning that AAs perform in social contexts. That is, we adopt the following protocol: whatever chains of reasoning are used by AAs successfully in social contexts (and for which we have a computational account), these are sufficient for that AA to be worthy of its own trust.

11. Explanatory and Interpretive Self-Knowledge

A human agent can have a variety of kinds of self-knowledge, such as interpretive and explanatory self-knowledge (see Velleman [17]) For instance, a human agent can have explanatory self-knowledge when they know why they are performing some task, such as washing the dishes (because their spouse asked them to wash the dishes). The same human agent doing the same task can have interpretive self-knowledge if they theorize about washing the dishes. Perhaps their interpretation of that action is the following: I am washing the dishes because my spouse asked me to and this shows that I am a person who is rationally responsive to the plans and wishes of their spouse.

However, when do AAs have explanatory and interpretive self-knowledge? If they are programmed to ask why-questions of their own actions, then they can have explanatory self-knowledge. Furthermore, if they are programmed to theorize about what it is they do (to interpret what it is they do), they can have interpretive self-knowledge. There is no reason in principle why AAs cannot have the same kinds of explanatory and interpretive self-knowledge that human agent can have. What is the importance of this point? It is important because it shows that what an AA can know about itself is not restricted to a subset of what a human agent can know about itself. If the human agent's self-conception is determined by, in part, the kinds of self-knowledge that it can have, then an AAs self-conception can be similar to that of a human agent at least in that respect. Thus, an argument that human agents must have a different kind of substantial self from that of an AA cannot be surmounted on the grounds that human agents do, and AAs do not, have explanatory and interpretive self-knowledge.

Moreover, having both kinds of self-knowledge may play a role in the reasoning that an agent does in determining whether they are worthy of their own trust. Here is an example of how such reasoning might proceed in a human agent. Suppose that a human agent surveys the reasoning he performs that involves principles of logic. He might ask of himself why it is that he is doing this survey. One explanation might be that he is doing the survey because he wants to make sure that he understands the principles of logic and that he is using them correctly. Additionally, he might theorize about what it is that he is doing. For instance, he might theorize about his surveying of the principles of logic in this particular context that it involves understanding the principles of logic and that this understanding will consist in being able to correctly use the principles and being able to see that he is correctly using the principles. The latter is a matter of being able to understand the principles and then apply them to specific problems. There is no reason why an AA could not come to have similar kinds of explanatory and interpretive self-knowledge. Thus, there is no reason why an AA could not use such kinds of self-knowledge in determining that it is worthy of its own trust.

12. The Varieties of Trust and the Existence of a Substantial Self

"Here it is tempting to load a great deal into an account of trust, and so to personalize it, or narrow it. Doing so can produce rich accounts of particular kinds and circumstances of trust, but it will fail to comprehend in a single account all of the varieties of trust there are" (Walker [8], p. 75). Is the existence of a substantial self—that which ensures that the stream of experiences that are the contents of our memories are unified—a feature of human beings that is not an essential feature of a trust relation, but, rather, is a feature that plays a role in certain kinds of trust, but not in others? Is the existence of a self merely part of a personalized account of trust and not something that can be comprehended in a single account of all of the different kinds of trust? If the answer to either of these questions is 'yes,' then the reductionist about the self can argue (i) that there is no need for a substantial notion of the self in accounts of trust, for there are some varieties of trust that do not require a substantial self (in which case, those varieties of trust that do require a substantial notion of the self can be deflated in some way) and (ii) it might even be the case that the only use of an substantial notion of the self is in personalizing accounts of trust.

On the reductionist view that there is no self, the idea of a substantial self will drop out of the conceptual framework of trust. However, things are not so simple. Even if reductionism about the self is true, human agents might not believe it and instead belief in the existence of a substantial self. Such a false view about the self would still be at work in the inferences and beliefs engaged in and held by a person who had such a false belief. Indeed, if everyone held such a false belief, then, even though reductionism about the self is true, people would act as if it is false. What would the difference be, if reductionism about the self is true, but people act as if it is false, and where reductionism about the self is true, and people do not have a false belief in the existence of a substantial self? The difference would be qualitatively significant. Here is what the difference would consist in. If someone believes in the existence of a substantial self, then they believe that even though there might be great changes in their psychology over a long temporal interval, that being is still them even after those changes have occurred. The changes have occurred to a substantial self, which unifies all of the psychological events over that temporal interval. Thus, a substantial self can survive great psychological change. Even if it turns out that there is no substantial self, believing that there is one will result in different behavior from not believing that there is one. Should the Buechner–Tavani model of trust be adjusted to take into account these possibilities?

We have the following four possibilities: (i) there is a substantial self and human agents believe there is one (ii) there is a substantial self, but human agents do not believe there is one (iii) there is no substantial self, but human agents believe there is one and (iv) there is no substantial self, and human agents do not believe there is one. We have already seen (above) that not believing there is a substantial self can undermine formation of normative expectations of one human agent toward another human agent. Does the Buechner–Tavani model of trust need to take into account such situations? Here we must guard against the following misunderstanding. Someone might object that the model of trust does not need to take such situations into account, since the model provides a conceptual elucidation of the notion of trust, and that elucidation should not depend upon the beliefs of human agents. For instance, a cognitive model of human problem solving should not take into account the mistaken beliefs human agents might have about how to solve problems. If those beliefs interfere with problem solving, it is not the point of the model to explain such interferences. A different model—one that explains how beliefs can interfere with the cognitive skills—would be the appropriate model to describe such situations (see Pylyshyn [18]).

But a conceptual model of trust is different from a cognitive model of some cognitive skill. It needs to have the resources to describe what happens in situations where there are factors which undermine some of the features of the model—such as the feature of normative expectation of one human agent toward another human agent. The model needs to have the resources to explain why it is that such undermining of features of the model can occur. However, the Buechner–Tavani model of trust does have the resources to explain such situations. For instance, suppose that as a result of having a belief in reductionism about the self, other human agents are unable to have normative expectations toward that human agent. If so, there cannot be a relation of trust between that human agent and all other human agents who cannot have a normative expectation toward that agent. The model can account for the failure of trust between those human agents.

There is one problem for the Buechner–Tavani [1] model of trust posed by reductionism about the self. Suppose that the reductionist about the self-claims that we should re-define trust (in the same way that the reductionist about the self has re-defined personal identity as satisfaction of relation) in order that human agents who are reductionists about the self can satisfy the relation of trust (in the way in which the reductionist about the self understands it). In that case, there would be two notions of trust (until either reductionism about the self is shown to be true or to be false), one for those who believe in reductionism about the self, and one for those who believe in a substantial self. How should this be settled? If reductionism about the self is true and we demonstrate that it is true, then we would have to opt for a re-definition of trust which did not advert to self-trust or to the pro-reactive pro-attitudes. Indeed, this re-definition of trust would look like those definitions of trust that defined it in terms of simple expectations. It would soon become evident that the re-definition suffers many counterexamples. At that stage, either the reductionist about the self would have to admit that trust between human agents is illusory or that trust based on simple expectations is all we can have for a trust relation.

13. Awareness of a Human Agent and Good Will toward Human Agents

What does awareness in a human agent consist of? And is it an essential feature of a trust relation? That is, will a certain kind of awareness have a normative role in trust? If the answers to these two questions are both 'Yes,' then there is a problem for reductionism about the self. Thus we need to ask whether there is a distinction between awareness of relation R and awareness of a substantial self. Certainly, the two are distinct in at least one significant way: the concepts are different, and so one is aware of something that has a conceptual component that differs in each case. However, being aware of the human agent next to me qua their substantial self and being aware of the person next to me qua relation R—how do these two kinds of awareness differ?

One might object that unless the solution to the problem of other minds is at hand, it is pointless to examine differences between these two relations. However, that objection is diminished by the fact that human agents do not in their everyday affairs worry about the existence of other minds, and it is in the context of our everyday affairs that we develop relations of trust toward other human agents.

The awareness of relation R would simply be awareness of some relation. However, awareness of a substantial self is not awareness of a relation, but of another human agent. Do these differences make for a difference in satisfying the conditions in the Buechner–Tavani model of trust? I believe that they do. If a human agent is not aware of himself—indeed, cannot be aware of himself because there is no substantial self of which to be aware—then conditions (vi) and (vii) in the model cannot be satisfied. Moreover, conditions (i)–(v) in the model cannot be satisfied if human agent A cannot be aware of human agent B in the sense that human agent B has a substantial self of which human agent A has an awareness (awareness—as well as unawareness—can be computationally implemented in AAs using modal operators for awareness (and for unawareness). If an AA has the concept of a substantial self in its lexicon, then that can serve as a predicate in the matrix of the modal operator).

What is it for one human agent to have an attitude of good will toward some other human agent? How is an attitude of good will toward them different from having a belief about them—the belief that one views them with good will? One difference is that the belief is just that—a statement of the information that one views them with good will. However, an attitude brings with it many different beliefs, as well as acts, capacities, and norms. It is not exhausted by a set of beliefs, no matter how informationally rich that set of beliefs happens to be, for there is also a matter of acts, capacities, and norms. A set of beliefs does not have normative status on its own unless it is augmented with an act or capacity, such as recognition of the normativity of some action (where the information that the action has normative status is expressed in a belief). Believing that the right thing to do is such-and-such in certain circumstances does not have a normative claim on anyone. It is only when one recognizes that the act is the right thing to do in those circumstances that one engages with its normativity, and it is only when one decides to do the action because it is the right thing to do that one is bound to its normativity.

However, if there is no substantial self, then one human agent can have beliefs about whether another human agent satisfies relation R, but it seems that they could not have a normative attitude—such as that of good will—toward the other human agent. What would that mean in the context of establishing a trust relation? Where a human agent cannot have a normative attitude of good will toward another human agent, it is unlikely that they would be able to trust the other human agent to act responsibly and to bear toward them the normative expectation that they will do such-and-such. Of course, it is not out of the question—it is conceptually coherent—that they could bear such attitudes even where they do not bear the normative attitude of good will toward another human agent. However, there are kinds of specific trust that require bearing good will toward the other human agent, and these kinds of specific trusting relations would not be available on a reductionist view about the self.

14. Can Expressivism Help the Reductionist about the Self to Establish Trust Relations?

An objection to the claim that a reductionist about the self cannot countenance normative attitudes of one human agent toward another human agent might be that on a naturalistic account—that is, an expressivist account—of normativity, all that matters for manifesting normative attitudes are naturalized features of a human being—such as their speech (see Gibbard [19–21]). Thus, there is no problem for a reductionist about the self to provide room for normative attitudes. In response to this objection, even if it is conceptually coherent that only naturalized features of a human being are necessary and sufficient for manifesting normative attitudes, it is still the case that the real emotions and real feelings that are partial causes of manifesting normative attitudes would have no bearer for the reductionist about the self—no substantial self that bears those real emotions and real feelings.

Instead, there would be relation R, which would be the locus of those real emotions and real feelings. However, how could the latter have a causal influence upon the former? After all, it is a series of memories that satisfy relation R, and how could a series of memory states be causally influenced by a certain emotional state? Can a series of memory states bear computational relations to one another and to emotional states? The burden of proof is on the reductionist about the self to show that there could be such computational relations between memory states and between memory states and emotional states.

15. First-order and Second-Order Beliefs and Reactive Attitudes

First-order and second-order reactive attitudes—and first-order and second-order beliefs—are important to distinguish, since each plays a role in the trust relation (see von Wright [22]). Second-order beliefs are about the content of beliefs of first-order reactive attitudes, and the second-order beliefs are themselves the objects of reactive attitudes—the second-order reactive attitudes. Both kinds of reactive attitudes have a normative force, while both kinds of beliefs have informational content, but no normative force. Second-order reactive attitudes arise in specific kinds of trust relations. Here is an example: John forms a first-order reactive attitude toward Jane of the following kind: he normatively expects of Jane that she will do such-and-such. Perhaps he knows Jane fairly well, and that he has trusted her in the past to do other things (than such-and-such). He might then form a second-order reactive attitude toward himself of the following kind: I normatively expect of myself that I will normatively expect of jane that she do such-and-such because I have known Jane fairly long and have trusted her in the past to do various things for me (other than such-and-such).

How does a second-order belief arise in the context of a trust relation? Suppose that Jack believes that he can trust Jill to do such-and-such. When he reflects upon this belief, the reflection can take many different forms. For instance, Jack might think that he is glad to have the belief that he can trust Jill. Of course, it would be better for Jack if he simply were glad that he can trust Jill—this need not be a second-order belief. Being glad to have a belief is different from having the belief that you are glad, even though each is a belief. The former is a second-order belief, while the latter is a first-order belief, and each has different belief contents.

The act of reflecting upon trust—which is a second-order belief about trust—reveals the commitments that trust requires because it itself is a commitment. Human agents can make a commitment to a first-order belief—can they also make a commitment to a first-order attitude? Yes—that is conceptually coherent and no doubt happens in specific kinds of trust relations. Here is an example. Jack forms a normative expectation that Jill will do such-and-such. He decides that he is committed to forming this normative expectation because of its importance to his life plans of the near future. There could also be a third-order reactive attitude and third-order beliefs in specific kinds of trust relations. There might, indeed, be an ascending hierarchy of orders of reactive attitudes and of beliefs.

The importance of nth-order reactive attitudes and beliefs in trust relations is the following: When human agents form second-order reactive attitudes to first-order reactive attitudes, the content of the second-order attitude is generally toward that human agent, while the content of the first-order reactive attitude is generally toward the other human agent (in the trust relation). However, where the human agent fails to believe that there is a substantial self, there will be difficulties in establishing the second-order reactive attitude, since the human agent might also believe that her future self will be so different from her current self that the future self will not satisfy relation R. If that happens, the agent will not form the second-order reactive attitude, and might then take back the first-order reactive attitude toward the other agent—she normatively expects the other agent to do such-and-such.

Notice that it would not be a difficult problem for AAs to have nth-order reactive attitudes and beliefs. However, where AAs believe that they have no substantial self (even if there is no computational simulation of a substantial self), problems will arise of the kind noticed in the preceding paragraph. Nth-order reactive attitudes and beliefs arise in specific kinds of trust relations. However, those who are reductionists about the self might not be able to engage in such specific forms of trust—this includes both human agents and AAs. This, then, is another example of a phenomenon—nth-order reactive attitudes and beliefs, in this case—that pose a difficulty for the trust relation for those who are reductionists about the self.

Second-order trust of oneself is a second-order reactive attitude toward oneself that one is responsible for being worthy of one's trust (the first-order reactive attitude). It is the second-order trust of oneself that implies that one will honor the commitments that arise as the trust relation is realized in a world in which there are changing circumstances and changing jobs, functions, duties, and obligations. It is this second-order trust of oneself that is jeopardized by reductionism about the self, since there might be no stable self that survives over an extended temporal period.

It is reflection on the trust relation with respect to the things that are left implicit but which must be addressed in performing the tasks that are part of the trust relation that one makes a commitment to ensuring that one will act responsibly and do what is best in that situation.

It is reflection on the trust relation with respect to the things that might go wrong in the local environment—things that would jeopardize doing the tasks that are part of trusting someone to do them—that one makes a commitment to ensuring that even if things do go wrong, that one will act responsibly and do what is best in that situation.

16. Autonomy, Trust and the Substantial Self

There is an important question concerning the relation between the idea of autonomy and the idea of a substantial self (see Frankfurt [23]). Without having a belief in a substantial self, can one believe that one's actions are autonomous? Can one believe that a human agent's actions are freely chosen if there is no substantial self? Let's consider the second question first. Prima facie, one can believe a human agent's actions are freely chosen—it is that human agent that satisfies the R-relation that freely chooses an action. However, we need to be careful—to not assume that a human agent has a substantial self. If the human agent has no substantial self, then just what person is it? Minimally, it is any person that satisfies relation R (and has those memories that satisfy relation R). Why couldn't such a person act—i.e., have the power of agency? Of course, that person that satisfies relation R might be a human agent, or it might be an AA with a set of memories (in which case, it is not a person). However, that is not what is important here. What is important is that if the person that freely makes the decisions is determined by relation R, and it is conceptually coherent that relation R could determine different persons at arbitrarily different times, it is not clear whether one human agent at time t_0 (who forms the intention to do such-and-such) is the same human agent at t_n (who fulfills the intention by an action). If the human agent can differ from t_0 to t_n, then it makes no sense to say that the first human agent acted freely, since he never acted, and it similarly makes no sense to say the second human agent acted freely, since he never formed an intention to act. Notice that the preceding considerations will hold of both human agents and AAs (that have programs sophisticated enough to simulate free will).

Let's now consider the first question that was raised above. Can whatever person that satisfies the R-relation be autonomous in the actions that they perform? If there is no owner of one's actions, then, by definition, whatever satisfies the R-relation cannot be autonomous. Furthermore, it is the substantial self which is usually taken to be the owner of one's actions. In giving up the substantial self, one also gives up the ownership view—that there is an owner of beliefs, actions, desires, wishes, and so on. For if there is no substantial self to be the owner of one's beliefs, actions, desires, wishes, and so on, then what else could possibly claim ownership of those things? The satisfier of the R-relation *could be a substantial self*—but Parfit proposes that the satisfier of the R-relation supersedes the substantial self. The satisfier is just the set of beliefs and memories that stand in the R-relation—and this set (of beliefs and memories) is hardly something that could be an owner of those beliefs and memories. It is a mere extension—the set of those beliefs and memories that satisfy the R-relation. There is no additional owner of those beliefs and memories. One specifies a human agent on a reductionist view of the self as just a set of beliefs and memories (and possibly desires). There is no deep further fact about a substantial self that unifies those objects in the set.

Under normal conditions, one can be aware of one's substantial self (contra, Hume [24]). That awareness grounds an intuition of autonomy of one's actions. Of course, if the substantial self does not exist, then one is aware of a non-existent object—in which case, one is having an awareness hallucination and it would follow that autonomy is then similarly an hallucination based on an awareness hallucination. However, a thorny question concerns the content of a belief state in which one believes in one's substantial self, but where the reductionist view about the self is correct—that there is no substantial self. Can believing that there is a substantial self thereby constitute a substantial self? Can believing that one is the owner of one's beliefs, actions, desires, wishes, and so on, *make* one the owner of them? Or are there necessary and sufficient conditions for the constitution of the self which are independent of any kind of belief about the self? If those necessary and sufficient conditions are satisfied, then one has a substantial self. If not, then one has no substantial self. However, believing that there is a substantial self, in that case, would not make it the case that there is a substantial self about which one has the belief. Instead, it would simply be that one's belief that one has a substantial self is true.

17. Self-Identification for Human Agents and for AAs

Do I need to have a substantial self in order to correctly self-identify myself (see Bermudez [25])? Can I self-identify myself even if there is no substantial self? If 'yes' is the answer to the latter question, then *what* is it that I am self-identifying? The answer for a reductionist about the self is easy: I am self-identifying me—and this me can be individuated purely in terms of the body I occupy and the connected memories that comprise my mental life (i.e., the satisfaction of relation R). There need be no substantial self for me to self-identify myself. In this way of speaking, 'myself' is just a synonym for my body and the connected memories that comprise my mental life. Additionally, 'me' picks out a unique body and unique mental life. I need not worry about misidentifying who I am on the basis of some false description.

But this view of self-identification by the reductionist about the self does not hold water. Here is a problem with it. I will be self-identifying on the basis of descriptions—a description of the body that I occupy and a description of the connected set of memories that comprise my mental life. However, unless I already know that those descriptions are truly descriptive of *me*, I can easily be mistaken as using those descriptions to make the self-identification. I might falsely identify some other person as myself. In the absence of a substantial self—which is *me*—I might erroneously self-identify. This is a serious problem for the reductionist about the self who believes that there can be an authentic trust relation between any two human agents (who satisfy the seven conditions in the Buechner–Tavani model of trust). For if the reductionist about the self can be mistaken about self-identification, then it is not clear that the seven conditions in the model can all be satisfied. For instance, conditions (vi) and (vii) might not be satisfiable for one who is a reductionist about the self, since the self that is the object of self-trust might be falsely identified. In which case, the human agent who makes the false identification will not satisfy conditions (vi) and (vii) in the Buechner–Tavani model.

One might think that the preceding problem for human agents who are reductionists about the self will not arise for AAs, even where AAs believe in reductionism about the self. After all—the objection goes—AAs cannot possibly have a substantial self. They can have memories (in the form of, say, linguistic descriptions) that are stored in a database. If these memories satisfy the R relation, then the AA is correctly identified. Thus, AAs could have self-trust.

Here is a simple way in which an AA can have self-trust: simply let the AA identify itself with some description—such as a description of a set if memories stored in its database. In this way, the AA knows which AA to trust—it turns out to be that AA itself. There is no need for the AA to have a concept of a substantial self in order to successfully identify itself as the agent whom it should trust.

Once the agent can successfully identify itself, it can also successfully refer to itself. In this way, there is no need to decide between reductionism about the self and a conception of a substantial self.

But there is a problem. It is this. The AA will not know that the AA it successfully identifies as itself is, indeed, itself. Without a concept of a substantial self, an AA can successfully identify itself without knowing that what it identified is itself. The question that arises is then: are there contexts in which the AA needs to know that the AA it successfully identifies is itself? If not, then no problems of making incorrect inferences concerning the AA will arise. If yes, then what are the incorrect inferences the AA might make about itself and how important is it that the inferences are correct?

One example is the following. Suppose that the AA is itself subject to a penalty if it fails to perform some informational task, that the penalty is waived if the AA does some other task, and that the AA is a member of a community of AAs. The AA successfully identifies an AA who must incur a penalty. That AA is itself. However, the AA does not make the identification of the AA as itself. So the AA does not perform the other task for which the penalty is waived. Suppose the community of AAs is engaged in some important task for human agents. The penalty that the AA must pay makes it impossible for the community to successfully perform the important task for human agents.

Could an AA successfully identify itself and know that the self it identifies is itself under a reductionist view of the self? Suppose that we stipulate that the 'memories' of the AA consist of its informational transactions over a certain time period. The question that arises is this: can an agent use those informational transactions to identify itself, or must it first know that it is the owner of those informational transactions in order to establish that they are the informational transactions that it has performed? Compare this with human agents. Does a human agent identify herself on the basis of the memories that she has or does she already have to know it is her before she can establish that those memories are her own memories? In both cases, if the latter option holds, then a substantial self is necessary in order to identify a connected sequence of memories. If the former holds, then no substantial self is necessary to identify the sequences of memories as being those of the self whose memories they are. As we saw above, a substantial self is necessary for a human agent to correctly self-identify herself, and the same is true for AAs—a substantial self is necessary for AA self-identification.

18. Summary

The Buechner–Tavani model of digital trust is revised—new conditions for self-trust are incorporated into the model. These new conditions raise several philosophical problems concerning the idea of a substantial self for social robotics, which are closely examined. In Sections 6 and 8–17 above, I have argued that there are insuperable difficulties for establishing a trust relation where there is a reductionist view of the self and that these difficulties do not arise for a substantial self. These sections argue about issues of trust for human agents and AAs with respect to reductionism about the self and the existence of a substantial self for the following: desires, motivation, self-trust, explanatory and interpretive self-knowledge, the varieties of trust, awareness of human agents, good will toward human agents, expressivism, first-order and second-order beliefs and reactive attitudes, autonomy, and self-identification. On the basis of these arguments, I conclude that reductionism about the self is incompatible with trust relations between human agents, between human agents and artificial agents, and between artificial agents, while the existence of a substantial self is compatible with such trust relations.

Funding: This research received no external funding.

Acknowledgments: I would like to thank Herman Tavani and Lloyd Carr for discussions that led to the formation of several ideas in this paper and for reading a preliminary manuscript. I would also like to thank the two anonymous reviewers for many valuable comments and important suggestions which helped to improve this paper.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Buechner, J.; Tavani, H. Trust and multi-agent systems: Applying the "diffuse, default model" of trust to experiments involving artificial agents. *Ethics Inf. Technol.* **2011**, *13*, 39–51.
2. Buechner, J.; Tavani, H. Autonomy and Trust in the Context of Artificial Agents. In *Evolutionary Robotics, Organic Computing, and Adaptive Ambience: Epistemological and Ethical Implications of Technomorphic Descriptions of Technologies*; Decker, M., Gutmann, M., Eds.; Verlag LIT: Berlin, Germany, 2015; pp. 29–51.
3. Buechner, J.; Simon, J.; Tavani, H. Re-Thinking Trust and Trustworthiness in Digital Environments. In *Ambiguous Technologies: Philosophical Issues, Practical Solutions, and Human Nature*; Buchanan, E., de Laat, P., Klucharich, J., Eds.; International Society for Ethics and Information Technology: Corfu, Greece, 2014; pp. 65–79.
4. Taddeo, M. Defining trust and e-trust: Old theories and new problems. *Int. J. Technol. Hum. Interact.* **2009**, *5*, 23–35.
5. Lehrer, K. *Self-Trust A Study of Reason, Knowledge, and Autonomy*; Oxford University Press: New York, NY, USA, 1997.
6. Wright, C. On Epistemic Entitlement (II): Welfare State Epistemology. In *Scepticism & Perceptual Justification*; Dodd, D., Zardini, E., Eds.; Oxford University Press: New York, NY, USA, 2014; pp. 213–247.
7. Carr, L. Self-Trust and Self-Confidence: Some Distinctions. Available online: https://www2.rivier.edu/faculty/lcarr/Self-trust%20and%20self-confidence%20-%20some%20distinctions.pdf (accessed on 14 January 2020).
8. Walker, M. *Moral Repair: Reconstructing Moral Relations after Wrongdoing*; Cambridge University Press: New York, NY, USA, 2006.
9. Rosenthal, D. Awareness and Identification of Self. In *Consciousness and the Self New Essays*; Liu, J., Perry, J., Eds.; Cambridge University Press: New York, NY, USA, 2012; pp. 22–50.
10. Parfit, D. *Reasons and Persons*; Oxford University Press: New York, NY, USA, 1984.
11. Locke, J. *Essay Concerning Human Understanding*; Oxford University Press: New York, NY, USA, 1975.
12. Buechner, J. Two New Philosophical Problems for Robo-Ethics. *Information* **2018**, *9*, 256, doi:10.3390/info9100256.
13. Buechner, J. Does Kripke's Argument against Functionalism Undermine the Standard View of What Computers Are? *Minds Mach.* **2018**, *28*, 491–513.
14. Unger, P. *Identity, Consciousness, & Value*; Oxford University Press: New York, NY, USA, 1990.
15. Velleman, D. *Self to Self: Selected Essays*; Cambridge University Press: New York, NY, USA, 2006.
16. Madell, G. *The Identity of the Self*; Edinburgh University Press: Edinburgh, UK, 1981.
17. Velleman, D. *Practical Reflection*; Princeton University Press: Princeton, NJ, USA, 1989.
18. Pylyshyn, Z. *Computation and Cognition: Toward a Foundation for Cognitive Science*; MIT Press: Cambridge, MA, USA, 1986.
19. Gibbard, A. *Wise Choices, Apt Feelings*; Oxford University Press: New York, NY, USA, 1990.
20. Gibbard, A. *Thinking How to Live*; Harvard University Press: Cambridge, MA, USA, 2009.
21. Gibbard, A. *Meaning and Normativity*; Oxford University Press: New York, NY, USA, 2013.
22. Von Wright, G. Norms of Higher Order. *Studia Log.* **1983**, *42*, 119–127.
23. Frankfurt, H. *Necessity, Volition, and Love*; Cambridge University Press: New York, NY, USA, 1998.
24. Hume, D. *Treatise Concerning Human Nature*; Oxford University Press: New York, NY, USA, 1978.
25. Bermudez, J. *The Paradox of Self-Consciousness*; MIT Press: Cambridge, MA, USA, 1998.