information

Article

Two New Philosophical Problems for Robo-Ethics

Jeff Buechner [1,2]

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

[2] Saul Kripke Center, The Graduate Center, City University of New York, New York, NY 10016, USA

Received: 27 August 2018; Accepted: 16 October 2018; Published: 18 October 2018

Abstract: The purpose of this paper is to describe two new philosophical problems for robo-ethics. When one considers the kinds of philosophical problems that arise in the emerging field of robo-ethics, one typically thinks of issues that concern agency, autonomy, rights, consciousness, warfare/military applications, employment and work, the impact for elder-care, and many others. All of these philosophical problems are well known. However, this paper describes two new philosophical problems for robo-ethics that have not been previously addressed in the literature. The author's view is that if these philosophical problems are not solved, some aspects of robo-ethics research and development will be challenged.

Keywords: robo-ethics; robo-agents; moral reasoning; computational models of moral reasoning; functionalism; reliability; Saul Kripke

1. Introduction

There are many difficult problems for robo-ethics, especially from the perspective of computational modeling, such as (i) how to computationally model human moral reasoning in robo-agents, and (ii) should human moral reasoning be computationally modeled in robo-agents? This paper indirectly addresses these problems by posing two new philosophical problems raised by Kripke's argument against functionalism extended to robo-agents which (i) show that the above problems in computational modeling need to be reformulated, and (ii) may show that established work in robo-ethics needs to be reformulated. These two new philosophical problems are neither an artifact of computational modeling of human moral reasoning nor something imposed extrinsically from the outside, such as cost factors in developing software that models human moral reasoning. Rather, they are to do with the physical computers that physically implement the abstract computations which constitute robo-agents and, in particular, robo-agents that are programmed to perform human moral reasoning. The paper will first describe Kripke's argument against functionalism extended to robo-agents [1], before describing the two new philosophical problems it raises for robo-ethics. It then examines several ethical and legal issues in robo-ethics that are a consequence of these two philosophical problems. It concludes that if Kripke's argument is sound, robo-ethics is challenged. (The rest of this paper will use the phrase 'Kripke's argument' as an abbreviation of 'Kripke's argument against functionalism extended to robo-agents.').

2. Kripke's Argument against Functionalism Extended to Robo-Agents

This paper will show how Kripke's argument against the view in the philosophy of mind known as 'functionalism' (i.e., the philosophical view that the human mind is a computational object of some kind) can be extended to robo-agents. Kripke's argument starts from the well-known consideration that physical computers can undergo breakdowns, but it would be a serious mistake to interpret the argument as saying that all physical computers can undergo breakdowns, and thus that we cannot know with certainty what it is they compute. Kripke's argument does not say that—it is not a simple skeptical argument. Secondly, the argument does not assume nor need as a premise the claim that robo-agents have minds.

2.1. Worries about Physical Computer Breakdowns

Consider the following: (i) a physical computer might physically break down; (ii) the program the physical computer executes might have a bug so that it cannot compute the function it has been shown to correctly compute when running normally; and (iii) the input data to a given physical computer might be corrupted. These are all worries to which anyone who relies on a physical computer to make physical computations can easily attest. For researchers in robo-ethics these worries are well-known. There are many techniques for testing software and hardware for breakdowns, bugs, and data corruption. The state of the art in testing software for bugs is mathematically sophisticated [2–4]. However, hardware is another matter. Even quite reliable hardware—established as such by sophisticated testing procedures—can break down.

In most cases, we cannot perceive from a computer that it has broken down. However, suppose that we are able to prove conclusively that the program used by a physical computer for simulating human moral reasoning is correct and that the input data to the physical computer used to physically compute instances of human moral reasoning is not corrupted. It does not follow that the physical computer is correctly simulating human moral reasoning, since it might not be operating under normal conditions (for correctly simulating human moral reasoning). If it has undergone a breakdown, then it might be physically computing something other than a simulation of human moral reasoning. Philosophers worry about the nature of physical computation. But what philosophical problem could there be for robo-ethics in the possibility of computer breakdown and malfunction?

2.2. The Argument

In what follows, Kripke's argument is presented as it applies to physical computers that execute programs for robo-agents. The basic idea of the argument is that we cannot, by observing its output behavior, acquire the knowledge that a robo-agent is (i) operating normally (for computing, say, moral view A); (ii) that it is correctly computing moral view A; (iii) that it is executing program P (for computing moral view A); and (iv) that it is using data structure D (used in computing moral view A) *unless we make a stipulation that it satisfies (i)–(iv).* The only currently available sources for Kripke's argument are Buechner [5,6], which overlap, in parts, this paper. Kripke's argument starts with the philosophical views that (i) physical computers physically realize abstract computational objects (such as algorithms and programs); (ii) engage in physical realizations of abstract computational processes—such as computing the values of a function using an algorithm which is executed in a program; and (iii) a computer is a physical machine which physically realizes the abstract diagram of some abstract function.

2.2.1. Physical Breakdown

Robo-agents can break down in various ways, and when they do, they might not physically compute, say, moral view A, which the robo-agent would correctly compute if it had operated normally (for computing moral view A) and did not break down. That robo-agents can break down and malfunction raises the question of when they have (and when they have not) physically realized the computation of moral view A. This appears to be an uninteresting skeptical question—breakdown of a physical machine is a physical possibility, but *any* physical machine (such as a washing machine) might break down. We do not infer from the possibility of a breakdown of some physical machine that we cannot determine (i) when it has broken down, (ii) that we do not know what it has done under breakdown, (iii) that we do not know what it would compute if it had not broken down unless we stipulate (i)–(iii). However, Kripke's argument is not the claim that a physical computer might break down, and thus not correctly compute moral view A.

A physical computer might break down in physically computing moral view A. However, when a robo-agent is functioning normally in the computation of moral view A (i.e., it is not undergoing a break down in the computation of moral view A), and the software for computing moral view A is correct, it will output moral view A when it is given as input, say, a request for moral view A. However, there might be another moral view, B, which the same robo-agent is computing when it is computing moral view A. How could that be if it is functioning normally (in the computation of moral view A) and the software for computing moral view A is correct? To understand Kripke's argument, one must understand how it is that *under the circumstances that the robo-agent is functioning normally (in the computation of moral view A) and executing correct (i.e., verified to be correct) software for computing moral view A, it is also computing moral view B*. Here is how that can happen. When the robo-agent is operating normally *in the computation of moral view A* it is undergoing a breakdown *in the computation of moral view B*. Similarly, when this same robo-agent is undergoing a breakdown *in the computation of moral view A*, it is operating normally *in the computation of moral view B*. What are normal circumstances for computing moral view A are breakdown (or abnormal) circumstances for computing moral view B. What are normal circumstances for computing moral view B are breakdown (or abnormal) circumstances for computing moral view A. That is how a robo-agent computes moral view B under breakdown when it is functioning normally in the computation of moral view A. It is not that either moral view A or moral view B is being computed (where 'or' is read as exclusive or). Rather, it is that the physical computer computes moral view A and moral view B at the same time (in the same sequence of computational states).

If you think this claim is implausible, consider:

FACT The physical conditions under which a physical computer physically computes moral view A (under normal conditions for computing moral view A) are the *very same* physical conditions under which the same physical computer (in the same computational states) physically computes moral view B (under breakdown conditions for computing moral view B).

This is an easily verifiable *objective fact* about a physical computer and it is an easily verifiable objective fact about physical computations. The following three Sections (Sections 2.2.2–2.2.4) are simple consequences of this objective fact about physical computers and about physical computations. If you resist these consequences of FACT (by, say, taking them to be implausible, absurd or unimportant), then you are implicitly denying what is an easily verifiable fact-of-the-matter about physical computations.

2.2.2. Epistemic Indistinguishability

It follows from FACT that there is no evidence by which we can disentangle computation of moral view A from computation of moral view B. Why should that be the case? The reason why this is the case is simple. There is no evidence for disentangling the normal conditions for computing moral view A from the breakdown conditions for computing moral view B because they are the same set of conditions. Similarly, there is no evidence for disentangling the normal conditions for computing moral view B from the breakdown conditions for computing moral view A because they are the same set of conditions. Computing moral view A (under normal conditions) is epistemically (and observationally) indistinguishable from computing moral view B (under breakdown conditions). *Any* evidence for the claim that the physical computer is computing moral view A is also evidence for the claim that it is computing moral view B, and conversely.

Suppose we examine the output behavior of some robo-agent which we take to be computing moral view A. Even if the output behavior of this robo-agent is an instance of moral view A, we cannot conclude that the robo-agent is computing moral view A. This output behavior is compatible with the robo-agent computing moral view A and it is also compatible with it computing moral view B. It is computing moral view A where it is operating normally in computing moral view A and it is computing moral view B where it is operating under breakdown conditions for computing moral view

B. But the normal conditions for computing moral view A are the same conditions as the breakdown conditions for computing moral view B.

If we already know it is operating normally for computing moral view A, we can conclude it is computing moral view A. If we know it is operating under breakdown conditions for computing moral view A, we can conclude it is computing moral view A. But where and how can we acquire such knowledge? If we know the robo-agent is computing moral view A, then we know it is operating normally for computing moral view A or under breakdown conditions for computing moral view A. Similarly, if we know it is computing moral view B, then we know it is operating normally for computing moral view B or under breakdown conditions for computing moral view B. There is a circle in the reasoning: to conclude the robo-agent is operating normally, one must know it is computing moral view A. To conclude it is computing moral view A, one must know it is operating normally (for computing moral view A) or under breakdown (for computing moral view A). The output behavior of the robo-agent cannot be used to conclude it is computing moral view A unless one already knows what moral view it is computing or unless one already knows that the conditions of operation are normal (for computing moral view A)—or that they are breakdown (for computing moral view A). This knowledge cannot be acquired from observing the output behavior of a robo-agent.

This is not a matter of the weak skepticism that a robo-agent might undergo a breakdown and not succeed in computing moral view A. In that weak skepticism, we can still know when a robo-agent is operating normally (in computing moral view A) and when it has undergone a breakdown (in computing moral view A). According to Kripke's argument, we do not have that knowledge—we do not know whether the robo-agent has undergone a breakdown (in computing moral view A) or is operating normally (in computing moral view A) *unless we already know which moral view it is computing*. That is the principal difference between Kripke's argument and the skeptical view that a robo-agent might break down (and thus not succeed in computing moral view A). In the case of weak skepticism, we have the information that the robo-agent has broken down (in computing moral view A)—which we can use to determine what moral view the robo-agent is physically computing. In the case of Kripke's argument, we do not have that information—we cannot know when the robo-agent has broken down (in computing moral view A) and we cannot know when it is operating normally (in computing moral view A)—unless we already know that it is physically computing moral view A (Buechner [6] argues that Kripke's argument is not an underdetermination argument and that it is not a triviality argument. See Buechner [7] for an extended discussion of triviality arguments).

The same problem arises when we ask which algorithm (implemented in a particular programming language) the robo-agent is physically executing in its physical computation of some moral view. If it is claimed that the algorithm for computing moral view A is physically implemented in the robo-agent (because, say, the programmer recalls that is the algorithm that was given as input to the robo-agent), how do we know that the robo-agent has not undergone a breakdown of some kind, as a result of which the algorithm is not one for computing moral view A, but rather an algorithm for computing, say, moral view B? What are normal conditions in implementing the algorithm for computing moral view A that are breakdown conditions in implementing the algorithm for computing moral view B, and what are breakdown conditions for implementing the algorithm for computing moral view A that are normal conditions for implementing the algorithm for computing moral view B. If we already know the robo-agent is computing moral view A, then we know that it is correctly implementing the algorithm for computing moral view A. Here, too, there is a circle in the reasoning: to conclude the robo-agent is executing the algorithm for computing moral view A, one must know it is computing moral view A. To conclude it is computing moral view A, one must know it is executing the algorithm for computing moral view A. The output behavior of the robo-agent cannot be used to conclude it is computing moral view A—or that it is computing moral view B, unless one already knows what moral view it is computing or unless one already knows that it is executing the algorithm for computing moral view A—or the algorithm for computing moral view B. This knowledge cannot be acquired from observing the output behavior of a robo-agent.

All of the usual sources of evidence as to what a robo-agent is doing (in its output behavior) are not available to us to justify our claim that the robo-agent is physically computing moral view A. We do not have as evidence for the claim the robo-agent is physically computing moral view A what it outputs. We do not have as evidence for the claim the robo-agent is physically computing moral view A that the robo-agent is operating normally for physically computing moral view A. We do not have as evidence for the claim the robo-agent is physically computing moral view A that it is implementing the algorithm for computing moral view A. We do not have any evidence—any checkpoints—for confirming a claim that the robo-agent is physically computing moral view A. That there are no sources of evidence for the claim that the robo-agent is computing moral view A (and likewise for moral view B) is equivalent to the claim that (i) the robo-agent is computing moral view A and (ii) the robo-agent is computing moral view B are epistemically indistinguishable. What follows from this?

2.2.3. Stipulation

Unfortunately, it follows that we cannot know the robo-agent is physically computing moral view A unless we *idealize* its behavior to that of physically computing moral view A (similarly, we cannot know the robo-agent is physically computing moral view B unless we idealize its behavior to that of physically computing moral view B). But to idealize its behavior to that of physically computing moral view A is to *stipulate* that it is computing moral view A. We must also stipulate that it is operating normally in computing moral view A (when we take it to be operating normally in computing moral view A) and we must stipulate that it is undergoing a breakdown in computing moral view A (when we take it to be undergoing a breakdown in computing moral view A). We must also stipulate that it is implementing in some programming language the algorithm for computing moral view A (when we take it to be computing moral view A). Any human user who computationally interacts with the robo-agent will need to engage in a speech act, such as: 'I stipulate that this robo-agent is physically computing moral view A'. Although one might think it is absurd to believe such a speech act needs to be made, that is only because common sense tells us that engaging in such speech acts is irrelevant to the behavior of a robo-agent. If Kripke's argument is sound, such a common sense view will be abandoned.

These stipulations need to remain in force throughout the operations of any robo-agent. At any given temporal stage in the computation of moral view A (or any other computation the robo-agent makes), the stipulation that the robo-agent is physically computing moral view A (or any other computation it makes) needs to remain in force. Even after the computation ends, and one can see (by observation) that the robo-agent outputs all of the features in the computation of moral view A, the stipulation that moral view A has been physically computed needs to remain in force. Observing that the nth feature is the nth feature in moral view A tells one nothing about whether the robo-agent physically computes moral view A.

2.2.4. No Matters-of-Fact

It would seem that there is a fact-of-the-matter as to which moral view a robo-agent computes—just as there must be a fact-of-the-matter as to which computation a physical computer makes when it makes a physical computation. But following Kripke's argument, that is not so. There are no facts-of-the-matter as to which computation a physical computer makes. There is no fact-of-the-matter as to which moral view a robo-agent computes. This is a simple consequence of the fact that the physical conditions under which a physical computer physically computes moral view A (under normal conditions for computing moral view A) are the very same physical conditions under which the same physical computer physically computes moral view B (under breakdown conditions for computing moral view B).

Whichever computation a physical computer makes when it performs a physical computation is determined by a stipulation made by a human user. Similarly, whichever moral view a robo-agent computes is determined by a stipulation made by a human user. Before any stipulations are made, there are no facts-of-the-matter with respect to what is computed. After a stipulation has been made, there is a fact-of-the matter as to which computation has been made. But it is a fact-of-the-matter only relative to a stipulation.

Thus, there are no objective facts about what is computed before a stipulation has been made, and so there is nothing concerning what is computed about which we can be said to know it with certainty, or with less than certainty. It is only after a stipulation has been made that there is a fact-of-the–matter that we can know, and we know it with certainty. Why is that? In making the stipulation as to which computation has been made by a physical computer, it is an a priori truth that the physical computer makes the computation that it has been stipulated to make. In the case of a robo-agent, there is no objective matter-of-fact as to which moral view it computes. But once a stipulation has been made as to which moral view it computes, we know with certainty which moral view it computes. The stipulation establishes an objective fact-of-the-matter as to which moral view a robo-agent computes *provided that everyone else in the community agrees with the stipulation*. If not, the stipulation establishes a matter-of-fact relative to a restricted subset of the community.

Since in the absence of making a stipulation as to which moral view a robo-agent computes, there is no fact-of-the-matter as to which moral view it computes, and it makes no sense to ask how likely it is that a robo-agent computes, say, moral view A. Moreover, once the stipulation has been made, the fact-of-the-matter is that only one moral view (say, moral view A) is physically computed by a given robo-agent. So it is certain that moral view A is physically computed under the stipulation. Recall that there are no sources of evidence—no checkpoints—available to establish a fact-of-the-matter as to which moral view the robo-agent computes.

Note carefully that a stipulation does not ensure a physical computer will not break down. Rather, the stipulation ensures that the physical computer is either computing moral view A under normal conditions for computing moral view A or computing moral view A under breakdown conditions for computing moral view A. It ensures that the physical computer is not computing moral view B under normal conditions for computing moral view B or moral view B under breakdown conditions for computing moral view B.

The distinction between the robo-agent being correct in which moral view it computes and our saying it is correct in which moral view it computes vanishes when the human user who computationally interacts in a social context with the robo-agent stipulates that the robo-agent computes, say, moral view A (by also stipulating that the robo-agent is functioning normally in the computation of moral view A and that it is executing the algorithm for physically computing moral view A). If likelihoods cannot be assigned to the outputs of a robo-agent in order to justify claims about which moral view it computes, can we say that a robo-agent reliably computes moral view A? Worse, if a human user has to stipulate that the robo-agent computes moral view A, can we say that the robo-agent reliably computes moral view A? Where we have no evidence sources for establishing which moral view a robo-agent physically computes, we have no basis for establishing claims about the reliability of that robo-agent in computing moral view A.

Since we cannot speak in these cases of 'correctness,' nor of 'reliability,' we have a form of relativism that is dubbed here 'robo-agent relativism.' For an account of why the relativist (about truth) cannot distinguish between being right and thinking he is right, see Hilary Putnam [8]. For more on robo-agent relativism, see Buechner [6]. The only way to establish a fact-of-the–matter as to which moral view a robo-agent computes is to make a stipulation, and so what a robo-agent computes is relative to a stipulation made by a human user. If two different human users make two different stipulations as to which moral view a robo-agent computes, there is no fact-of-the-matter as to which stipulation is correct. Relative to one user, their stipulation is correct. Relative to another user, their stipulation is correct. Matters-of-fact about what robo-agents compute are, then, relative to those human users who make stipulations as to what robo-agents compute.

If your view is that Kripke's argument is unimportant or uninteresting for robo-ethics because it is an unimportant or uninteresting form of skepticism about physical computations, consider that even forms of skepticism which most of us take to be implausible (such as that we are now brains-in-a-vat) reveal interesting truths we would not have known had we not pursued the matter. (For an example, see Putnam [8], which established a connection between the Löwenheim–Skolem theorem and brain-in-a-vat skepticism.) Whatever we think of the two new philosophical problems for robo-ethics created by Kripke's argument, we can learn much from taking them seriously and responding to them.

3. Two New Philosophical Problems that Kripke's Argument Creates for Robo-Ethics

Prior to Kripke's argument, robo-ethics examined questions such as the following:

1 Should robo-agents perform risky surgery on human beings?

A consequence of Kripke's argument is that such questions need to be reformulated as follows:

2 Should robo-agents perform risky surgery on human beings given such-and-such stipulations as to what actions they perform?'

Question 1 is different to question 2. If Kripke's argument is sound, this difference between 1 and 2 must be addressed by many people working within the field of robo-ethics. That is to say: many questions arising in robo-ethics, in addition to 2, will need to be reformulated. The first philosophical problem for robo-ethics is whether 2 can be paraphrased by 1. That is, can the phrase 'given such-and-such stipulations as to what actions they perform' be eliminated without loss of meaning? Or is 2 qualitatively distinct from 1—in the sense that an answer to 2 need not be an answer to 1 (because the reasons for the answer to 2 are different reasons than the reasons for the answer to 1)?

The second philosophical problem for robo-ethics depends upon the first. If no paraphrase of 2 by 1 can be provided, the question then is how to understand 2—in particular, how to specify its truth-conditions. What will constitute acceptable reasons for an answer to 2? If many (most? all?) questions in robo-ethics need to be reformulated by adding the phrase 'given such-and-such stipulations as to what actions they perform,' then a good deal of research in robo-ethics will depend upon how this philosophical problem is treated.

Notice that we do not have such a problem of reformulation when we raise moral questions about what human beings do (and there is no interaction with robo-agents). For instance, we could raise a moral question about human surgeons:

3 Should human surgeons perform risky surgery on human beings?

But we do not raise the question:

4 Should human surgeons perform risky surgery on human beings given such-and-such stipulations as to what actions they perform?'

(It should be noted that these new philosophical problems for robo-agents will also arise for human beings only if functionalism (the view that human minds are computers) is taken as the correct theory of human mental states and mental properties. Reject functionalism and the problems disappear. However, functionalism (applied to physical computers) is part of the standard view of what computers are, and can hardly be rejected. See Buechner [6] for an elaboration of the remarks in this paragraph, and for more on different views about physical computers, such as enactivism.)

One interest in robo-ethics is whether robo-agents reliably simulate human moral reasoning. If the moral reasoning humans engage in to answer 1 is the same moral reasoning robo-agents engage in to answer 1, then robo-agents simulate human moral reasoning (in that context). If all of the instances of robo-agent reasoning in that context simulate human moral reasoning in that context, then the robo-agent moral reasoning reliably simulates human moral reasoning. The same can be said of 3. This would show that robo-agents can reliably provide moral reasoning about robo-agents and moral reasoning about human beings. Such results would be quite important in the field of robo-ethics.

However, if Kripke's argument is sound, and 2 cannot be paraphrased by 1, those results would be undermined. Since 2 is a different question than 1, the moral reasoning required to answer 2 might be different from that required to answer 1. There would be no such problem for human beings providing moral reasoning about other human beings (unless the question concerns human beings providing moral reasoning about how human beings should morally reason about 2). Human beings providing moral reasoning to answer 3 would do so differently to robo-agents, since whatever moral reasoning the robo-agent provides would require a stipulation by some human being that it is that moral reasoning and not something else (such as a different piece of moral reasoning).

Notice also that reliability is an epistemic notion and, to be a coherent notion, it requires matters-of-fact about which we can be right or wrong. Do robo-agents reliably simulate human moral reasoning? Since there is no matter-of-fact as to what a robo-agent computes in the absence of a stipulation as to what it computes, there is no matter-of-fact about which we can be right or wrong (other than what has been stipulated to hold of a robo-agent). It is not coherent to ask of a robo-agent whether it reliably simulates human moral reasoning. We cannot apply the epistemic evaluative term 'reliable' to actions of robo-agents unless we solve the philosophical problem of whether 2 can be paraphrased by 1.

Whether there are moral facts is an unresolved philosophical problem in meta-ethics. However, even if there are moral facts, we cannot evaluate the moral correctness of a robo-agent's moral actions before a stipulation has been made as to what moral actions it has performed. Indeed, no matter what meta-ethical view one has of moral values, one cannot determine whether the moral actions of a robo-agent accord with that meta-ethical view before a stipulation has been made. Thus, on the basis of stipulations for which there is no evidence as to whether they are or are not correct (and so we cannot apply the notion of 'being correct' to those stipulations), a moral decision must be made—such as whether a robo-agent should perform surgery (we can be correct about whether a stipulation has been made and correct about what stipulation has been made. But we cannot apply the term 'correct' to what has been stipulated). Ground-breaking work in meta-ethics (such as a successful argument that there are moral facts) might not be applicable to robo-ethics if 2 cannot be paraphrased by 1.

4. Some Possible Objections

There are several possible objections that a careful reader of this paper could raise. This paper considers nine objections, and argues that each fails either to refute its target or shift the burden of proof upon the present author.

4.1. Moral Values and Computation

One objection to the use of Kripke's argument is that moral values are not—as far as we know—computational objects. Moreover, the objection continues, reasoning about and with moral values occurs in a natural language and is not subject to the problem of which moral view is being computed and when conditions of operation for the computation of that moral view are normal. In sum, Kripke's argument focuses on the wrong target, and does not imply there are new philosophical problems for robo-ethics.

This objection might seem convincing until one realizes that whichever actions a robo-agent performs are computational (or the result of computations). Moreover, there are translations of what appear to be non-computational objects in high-level programming languages into computational objects. Every statement in whatever high-level programming language is used to program robo-agents will eventually be translated into sequences of 0s and 1s. If high-level constructs in a high-level programming language cannot be translated into sequences of 0s and 1s, they cannot be computed within a computational device. Moreover, if natural language sentences cannot be translated into a high-level programming language, they cannot be computationally expressed in a robo-agent.

4.2. Breakdown Conditions Might Result in Arbitrary Computational States

When a robo-agent breaks down and malfunctions, it might compute almost anything. Why think that if under normal conditions it computes moral view A, it will compute moral view B under breakdown conditions? If so, then Kripke's argument loses much of its force. After all, why should anyone worry that under normal conditions (for computing moral view A) a robo-agent computes moral view A which are also the breakdown conditions for being in some arbitrary computational state? Similarly, why worry that under breakdown conditions (for computing moral view A), a robo-agent is in some arbitrary computational state (not part of computing moral view A) for which those conditions are normal?

There are two ways of responding to this objection. The first is that even if a robo-agent computes moral view A under normal conditions and some arbitrary computational object (other than moral view B) under breakdown conditions, it is still true that the robo-agent computes moral view A (under normal conditions) or some arbitrary computational object (under breakdown conditions) and moral view A (under breakdown conditions) or some arbitrary computational object (under normal conditions). There are no facts-of-the-matter to decide whether the robo-agent is computing moral view A or some arbitrary computational object. Thus a stipulation—based on no evidence—must be made as to what the robo-agent computes, and so both of the new philosophical problems for robo-ethics are in play.

Indeed, one can say of the arbitrary computational state that it is a 'we-know-not-what' computational state, in which case one can say that a robo-agent is either computing moral view A or some 'we-know-not-what' computational state, and that there is no evidence to choose between them. Thus, it must be stipulated that the robo-agent is computing moral view A.

The second point: suppose that A represents moral view A and that B represents moral view B. Let C be some arbitrary computational behavior. Thus an output of A (by some robo-agent) happens under normal conditions for A, breakdown conditions for B, and breakdown conditions for C. An output of B is breakdown conditions for A, normal conditions for B, and breakdown conditions for C. An output of C is normal conditions for C, breakdown conditions for B, and breakdown conditions for A. The space of possible behaviors of the robo-agent has expanded—and the robo-agent must be idealized, stipulated—as computing either A, B, or C.

This second point has less force than the first point, since it uses a logical possibility to expand the space of possible behaviors of the robo-agent. The robo-agent might never output B—all that matters is that it is logically possible that the robo-agent could output B.

4.3. How Could a Physical Computer Compute Two Different Functions When There Is Only One Physical Process?

It is a mistake to think that the physical computer is:

(A) either operating normally or not—that is, either computing F or in breakdown (and so not computing F).

Rather, it is a fact about physical computations that:

(B) either (i) the computer is operating normally in the computation of F and operating under breakdown in the computation of G or (ii) operating under breakdown in the computation of F and operating normally in the computation of G.

For most people (including computer professionals) who have had any kind of exposure to physical computers, A is a deeply ingrained view. But A is false. B is true. B is true because there is only one physical process when the physical computer is operating normally in the computation of F and operating under breakdown in the computation of G and there is only one physical process when the physical computer is operating normally in the computation of G and operating under breakdown in the computation of F.

4.4. Surely There Is a Diagnostic Tool That Would Show the Physical Computer Is Computing Function F and That It Is Not Computing Function G

The objection is that there must be diagnostic tools which can reliably establish that the physical computer executing a certain algorithm (for computing function F) is physically computing F and is not physically computing function G. I make three responses to this objection. First, any software diagnostic tools must be implemented in a physical computer, and are thus subject to the problem—are they computing function A (under normal conditions) or are they computing function B (under breakdown conditions)? Second, even if the diagnostic tools confirm that function F is computed by a physical computer executing some algorithm (for computing F), that does not confirm that physical computer is not computing function G. *No* diagnostic tool could confirm this, since whenever function F is computed (under normal conditions for computing F), function G is also being computed (under breakdown conditions for computing G). Third, even if the physical computer computes function F on 10,000,000 different runs, that does not show it is not computing function G. The runs data could not show that, since whenever the physical computer computes F (under normal conditions), it is also computing G (under breakdown conditions). The output of the physical computer which appears to confirm it is computing F (under normal conditions) is the same output that occurs for computing G (under breakdown).

4.5. 'Unreliability' When Applied to Robo-Ethics Is a 'Category Mistake'

Suppose that 2 can be paraphrased by 1. In that case we can speak of the reliability of a robo-agents moral reasoning simulating a human being's moral reasoning. However, one might object that the use of the term 'reliable,' applied to robo-ethics is (what Gilbert Ryle [9] and other philosophers call) a *category mistake*. Moral judgments are not reliable or unreliable, period. However, 'reliable' is not used in that way in this paper. Rather, it is applied to whether the moral reasoning of robo-agents (i) simulates human moral reasoning and (ii) whether that moral reasoning will be duplicated time and time again when the robo-agent is in morally equivalent situations. Notice that a robo-agent might satisfy (i) and fail to satisfy (ii) or fail to satisfy (i) and satisfy (ii). In either case, the robo-agent would be considered unreliable—and it would not be a category mistake to make that judgment.

4.6. There Are No Philosophical Problems for Robo-Ethics When the Software for Robo-Agents Has Been Proven to Be Correct

The objection is that if the software for robo-agents has been proven to be correct, there are no philosophical problems for robo-ethics. The first response to this objection is that Kripke's argument concerns physical implementation of software in a physical computer. It says nothing about software which is not physically implemented. Suppose that software for robo-agents has been proven mathematically to be correct—say the software models human moral reasoning. No stipulation has to be made as to what the software does. Once that software is physically implemented in a physical computer, however, the physical computer might break down. If so, the robo-agent will not straightforwardly compute what it has been programmed to compute. There is no evidence to disentangle epistemically what it has been programmed to compute and it computing something else.

The second response is that it might be the case—as will be suggested immediately below—that designing software for robo-agents in light of Kripke's argument requires incorporation into the software features which would not be incorporated if Kripke's argument were either refuted or ignored. If 2 cannot be paraphrased as 1, then, for instance, robo-agents engaging in moral reasoning about other robo-agents will have to take into account that the actions of a robo-agent are what they are because of a stipulation (by a human agent). To do that, the software for the robo-agent must take this into account (how it does it will depend upon how the second philosophical problem for robo-agent is resolved).

4.7. The 'There Is Nothing Special about a Stipulation' Objection

There is nothing special about human beings having to stipulate that all of the actions attributed to a robo-agent have actually occurred. For an analogy, consider that in machine learning, what a robo-agent successfully learns might not be something humans can reliably know. A robo-agent trained by a data-set to recognize one kind of object might actually recognize another kind of object, because of, say, leakage in the data mining set (see [10] for a discussion of data leakage and [11,12] for examples of data leakage). There might be no epistemic means by which we can disentangle one from the other—no way in which we can reliably determine what it is that the robo-agent has actually learned. However, this is not an impediment to machine learning—provided that the robo-agent is successful in, say, detecting what we take it to have learned. Similarly, that we cannot determine what the actions of a robo-agent are without human stipulation is nothing special, so long as the robo-agent acts in ways that we expect it to act (on the basis of the software designed for it and without making any stipulations).

This objection misses the point. In many connectionist algorithms used in machine learning, it is not transparent what it is that the robo-agent learns because of the architectural features of the learning space [13]. However, Kripke's argument is not about that problem. Rather, it is the problem that is created by the possibility that the physical implementation of the algorithm in a physical computer might break down and malfunction. It would be senseless to stipulate that a robo-agent executing a connectionist learning algorithm learns A rather than B because we do not know, based on features of the connectionist learning algorithm, whether it learns A rather than B. But it would not be senseless to stipulate that a robo-agent learns A rather than B when what are normal conditions for learning A are breakdown conditions for learning B and what are breakdown conditions for learning A are normal conditions for learning B. This is not a matter of not knowing what it is that a robo-agent does because of the architectural features of the learning space. We can know with certainty that the algorithm the robo-agent implements is an algorithm for learning A (and not for learning B) because we have a mathematical proof that the algorithm is correct. But once that algorithm is implemented in a physical computer, we lose that knowledge, since we cannot know (with any degree of certainty) what it is that the robo-agent learns without making a stipulation. The 'there is nothing special' objection confuses one problem with another problem. The two problems are different, and they need to be treated differently.

4.8. So What If a Computer Might Break Down? That Has No Importance in Itself and No Importance for Robo-Ethics

The objection is that the breakdown of a computer has no philosophical interest because it is something that is routine and when it happens, we adjust for it by either by fixing the physical computer or discarding it. This objection ignores or implicitly denies the easily verifiable objective fact about physical computations which is the basis for Kripke's argument—the physical conditions under which a physical computer physically computes function F (under normal conditions for computing F) are the very same physical conditions for computing function G (under breakdown conditions for computing function G).

A consequence of FACT (see Section 2.2.1 above) is that a stipulation has to be made as to whether the physical computer computes function F or function G. Once a stipulation has been made, it is known a priori with certainty which function the physical computer computes. The imposition of a stipulation on physical computations of a physical computer raises two philosophical problems concerning statements of the form: physical computer P physically computes function F only when a human agent stipulates that it computes function F.

4.9. Kripke's Argument Is on a Par with Skepticism about Dreaming

The objection is that Kripke's argument is on a par with skepticism about dreaming. If one takes skepticism about dreaming to be uninteresting and implausible, then one will take Kripke's argument to be uninteresting and implausible. The response to this objection is that Kripke's argument and skepticism about dreaming are not on a par—they are quite different arguments. For one, there is a fact-of-the-matter as to whether I am (now) awake or dreaming. However, there is no fact-of-the-matter as to whether a physical computer is now computing F (under normal conditions for computing F) or G (under breakdown conditions for computing G), since whenever it is computing F (under normal conditions for computing F), it is also computing G (under breakdown conditions for computing G). Second, making a stipulation that I am now awake would hardly settle the question of whether I am (now) awake or dreaming. But a stipulation does settle the matter of whether a physical computer is (now) computing F (under normal conditions for computing F) or computing G (under breakdown conditions for computing G). Since the two arguments are not on a par, it would be a fallacy of reasoning to conclude from the claim (which many epistemologists would deny) that skepticism about dreaming is uninteresting and implausible that Kripke's argument is uninteresting and implausible.

5. Which Aspects of Robo-Ethics Are Targeted by Kripke's Argument?

This paper will examine work from various aspects of robo-ethics to show how the philosophical problems arising from Kripke's argument applies to them.

5.1. McCarty's 'Deep' Conceptual Models

L. Thorne McCarty [13] has argued that a key feature of an intelligent legal information system is a deep conceptual model—which was, in part, a reaction against the rule-based systems that were pervasive in artificial intelligence (AI) applications in the mid-to-late 1970s. What is a deep conceptual model and how does it differ from a rule-based system? McCarty uses the example of MYCIN [14] to illustrate what he means by a shallow rule-based system. MYCIN contains a large number of rules which can be used by a medical doctor to make judgments about the cause of a given set of symptoms which she observes in a patient. For a given set of symptoms, the rules determine the likely bacterial infection which causes those symptoms. MYCIN is shallow because it does not contain any representation of the causal mechanism by which the bacterial infection produces a set of given symptoms in a patient. Because there are no such representations in MYCIN, the system is unable to reason about the cause for a given set of symptoms. This is a shortcoming in cases in which there is causal overdetermination of a given set of symptoms, causal underdetermination of a given set of symptoms, and where the likelihood that a given set of symptoms is caused by some bacterial infection is quite low. In all such cases, one needs to reason about the causal mechanisms which produce the symptoms. Any shallow rule-based system is incapable of performing such reasoning. Deep rule-based systems are called 'model-based systems,' An example is CASNET [15], which contains a representation of many different kinds of relations and properties that occur in cases of glaucoma. For example, there are many different kinds of temporal relations, causal relations, and hierarchical relations between the different physiological states that have been observed to exist in cases of glaucoma. McCarty's aim in discussing MYCIN and CASNET is to reveal the distinction between shallow rule-based systems and conceptual model systems so that he can illustrate the usefulness of the latter in legal information systems, such as the TAXMAN system that he created in collaboration with N. S. Sridharan [16]. In TAXMAN several examples of a deep conceptual model are defined—such as the concepts TRANS and D-REORGANIZATION. These definitions consist of several software clauses. In physical computers executing TAXMAN breakdowns can occur. What are normal conditions for computing TRANS could be breakdown conditions for computing something else (such as NOT-TRANS). What are normal conditions for computing something else (such as NOT-TRANS) can be breakdown conditions for computing TRANS.

A deep conceptual model of a moral system, such as consequentialism, might consist of several concepts which are given explicit definitions in the software, such as CONSEQUENCES and UTILITY. Just as in TAXMAN, the software definitions of these concepts might be altered in a physical computer which undergoes breakdown. *Any* part of the software definition of a concept might be altered under physical computer breakdown, as might any part of the program in which those concepts are defined and used. Thus, Kripke's argument targets deep conceptual models employed in robo-ethics. Any context in which deep conceptual models are employed for some purpose will also be targeted. This is a powerful claim—deep conceptual models are the bedrock of robo-agent software designed for almost any purpose whatsoever within robo-ethics. To see how powerful this claim is, download Michael Anderson's moral reasoning system GenEth [17] and inspect the code. Each line of code in GenEth is subject to Kripke's argument. No line of code in GenEth can remain as is until the two philosophical problems for robo-ethics have been resolved (under the assumption that Kripke's argument is sound).

5.2. Tavani on the 'Moral Consideration Question' in Robo-Ethics

In a paper also included in this special issue of *Information*, Herman Tavani [18] has argued that one category of robo-agents, viz., *social robots*, should be accorded at least some degree of moral consideration. (Some might be inclined to interpret a claim of this sort as a necessary condition for robo-agents also being accorded rights; however, Tavani argues only that we should grant moral consideration to qualifying social robots as an alternative to granting them full-fledged rights because, for one thing, it enables us to circumvent many of the controversies surrounding "rights discourse" and the language of rights.) Suppose that there is a successful argument that robo-agents should be accorded moral consideration (by human agents) and that Kripke's argument is sound. No matter what those conditions are, it follows from Kripke's argument that whether they are satisfied or not by any given robo-agent must be stipulated by a human user of the software which implements those conditions—or a human agent who computationally and socially interacts with robo-agents. If there is no stipulation by a human agent that robo-agents satisfy those conditions, then there is no fact-of-the-matter as to what conditions they do satisfy.

We do not have:

5 Robo-agents satisfy such-and-such conditions. Therefore, they should be accorded moral consideration.

Instead, we have:

6 Robo-agents satisfy such-and-such conditions because a human agent has stipulated that they satisfy those conditions. Therefore, they should be accorded moral consideration.

Unless 6 can be paraphrased as 5, 6 is a different argument to 5. Even if 5 is sound, it does not follow that 6 is sound. This is a significant result, for it shows that arguments in robo-ethics about important issues—such as whether robo-agents should be accorded moral consideration—are also targeted by Kripke's argument. These arguments might play no role in the design of software for robo-agents—yet they are still targeted by Kripke's argument.

Consider a human agent, Jones, about whom there is a question as to whether he is worthy of moral consideration. Jones appears to satisfy all of the criteria under which a human agent is worthy of moral consideration. However, Smith presents a good argument which claims that Jones satisfies those criteria only if some human agent (other than Jones) stipulates that Jones does satisfy the criteria. Although such a case appears to be absurd, if it did arise we would surely say that it would be best to withhold judgment as to whether Jones is worthy of moral consideration, since the fact that another human agent needs to stipulate that Jones satisfies the conditions for being worthy of moral consideration could undermine any judgment that Jones satisfies those conditions. Notice that such a stipulation is not the same as the verdict of an expert that, for example, Jones is capable of, say, reflection upon his actions. Such an expert would administer tests to Jones on the basis of which she might conclude that Jones is able to reflect upon his actions. Suppose that in the case of making a stipulation that Jones satisfies the conditions, (i) no tests are administered to Jones—no information is obtained about Jones, and (ii) the stipulation is not made on the basis of any evidence that Jones satisfies those conditions. Given that there is no evidence on which the stipulation is based, one would be rational to withhold belief that Jones satisfies the conditions. Why should one not also adopt this view when it is a robo-agent in a similar set of considerations? The answer is that the second new philosophical problem for robo-ethics must be resolved before such claims can be properly evaluated.

A different argument for the claim that robo-agents should be accorded moral consideration is now considered. Tavani examines Hans Jonas's ethical framework in the context of questions about whether robo-agents that socially interact with human agents can qualify for moral consideration. Applying Jonas's framework, Tavani then notes that under certain conditions, robo-agents socially interacting with human agents can significantly enhance our ability, as humans, to act in the world. [19]. Suppose that it is true that robo-agents can enhance our ability to act in the world when software is functioning normally and that under breakdown that software fails to do so. Consider a simple example: a robo-agent can provide instructions to a human agent on how to navigate a new city. This can enhance the human agent's ability to act. Suppose the human agent has arrived in a city with which she is unfamiliar, in order to give a talk at an important business meeting. She socially interacts with a robo-agent that tells her how to get from her hotel to the meeting place. But if she has not stipulated that the robo-agent provides correct directions (under normal conditions for providing correct directions) and the robo-agent suffers a breakdown, providing incorrect directions, then she will not have an excuse if she is late to the meeting, since in the absence of a stipulation the robo-agent could be taken to providing incorrect directions (under normal conditions for providing incorrect directions which are also breakdown conditions for providing correct directions).

We do not have:

7 Robo-agents socially interacting with human agents can enhance our ability, as humans, to act in the world. Therefore, robo-agents should be accorded moral consideration.

Instead, we have:

8 Robo-agents socially interacting with human agents—where what the robo-agents do is the result of a human stipulation—can enhance our ability, as humans, to act in the world. Therefore, robo-agents should be accorded moral consideration.

Unless 8 can be paraphrased as 7, 8 is a different argument from 7. Even if 7 is sound, it does not follow that 8 is sound.

5.3. Property versus Relational Views in Robo-Ethics

Whether one holds a property view or a relational view of conditions that need to be satisfied by a robo-agent to be worthy of moral consideration, Kripke's argument presents difficulties for each [20,21]. On a property view, suppose that one property is consciousness. To describe the difficulty, there is no need to have a philosophically acceptable definition of consciousness. Suppose consciousness is simply defined as self-reflection. When a program for a robo-agent that implements consciousness is executed under normal conditions, the robo-agent has the ability to self-reflect. But those normal conditions for self-reflection might be breakdown conditions for storing some piece of data in memory. What are breakdown conditions for self-reflection might be normal conditions for storing some piece of data in memory. There is no way to tell whether the program is self-reflecting or storing data without stipulating that it doing one, but not the other.

On a relational view, suppose that a robo-agent behaves in relation to human agents as if it is conscious. Indeed, suppose that all of the social interactions between robo-agents and human agents have this feature—that the robo-agents behave as if they are conscious. However, the behavior of the robo-agent that leads the human agent to conclude that it is conscious must be stipulated by a human agent.

We do not have:

9 When robo-agents engage in such-and-such behavior in social interactions with human beings, those human agents infer that they are conscious.

Instead, we have:

10 When robo-agents engage in such-and-such behavior stipulated by human agents to hold of them in social interactions with human agents, those human agents infer that they are conscious.

Unless 10 can be paraphrased as 9, 10 is a different claim than 9. Whether one holds a property view or a relational view of robo-agents features, Kripke's argument targets each.

6. Who Is Legally and/or Morally Responsible for the Actions of a Robo-Agent?

There are various legal and moral issues that arise where human agents stipulate all of the computations that are made by robo-agents. One such issue is: who is responsible for the actions of a robo-agent? Suppose that there is a successful computational model of human moral reasoning and judgment—it attains the highest level of James Moor's [22] typology of moral agency for robots, i.e., 'full ethical agents'. However, different human users can idealize robo-software in different ways. Human user A might stipulate a robo-agent as achieving the highest level in Moor's typology, while human user B might stipulate the same robo-agent as achieving the lowest level in Moor's typology. Is A responsible for what the robo-agent does? Is B responsible for what the robo-agent does? Are both A and B responsible? Are neither responsible? How do we individuate robo-agent actions? For instance, can we say of some robo-agent that it successfully performed a moral action of a specific kind when different human users of the robo-software determining which actions the robo-agent performs might take that robo-agent to be performing different moral actions? Whose stipulation that the robo-agent performed a moral action of some kind counts? Will those human agents whose stipulations count be responsible (or share responsibility) for the actions of a robo-agent? I will examine three features of the concept of responsibility in light of Kripke's argument.

6.1. Liability Responsibility

A human agent has the capacity for liability responsibility when she satisfies certain criteria concerning intention, knowledge, being reckless toward the consequences of an action, and several more [23]. When a human agent stipulates that a robo-agent performs such-and-such actions, does the human agent intend and know that the robo-agent performs those actions? It would appear that the answer is 'yes'. The same human agent might be reckless toward the consequences of the robo-agents actions. If a corporation selling the robo-agent software mandates that the human agent make a particular stipulation, the choice of a stipulation is thereby inevitable. Thus, human agents who stipulate the actions of a robo-agent have the capacity for liability responsibility.

The Problem of Lying to Avoid Non-Exculpatory Moral Loss

Suppose one human user who computationally and socially interacts with a robo-agent stipulates it as computing moral view A, while another human user stipulates it as computing moral view B. For the first human user, that robo-agent uses moral view A as a reason to perform certain actions that have causal effects on human agents. Suppose further that some of the causal effects are harmful because the program implementing moral view A in the robo-agent is defective ('Defective' can mean several different things—such as (i) is wrong about the concepts underlying moral view A, or (ii) is right about the concepts underlying moral view A, but does not see that application of those concepts in a real-world situation of a certain kind is inappropriate. This paper assumes the second interpretation of 'defective' for this example). Under the stipulation that the robo-agent has computed moral view A, there is a non-exculpatory moral loss, while under the stipulation that it computes moral view B, there is an exculpatory moral loss [24]. The human user who stipulates that the robo-agent computes moral view A bears liability responsibility and is to blame for the actions of that robo-agent. If it is moral view B that is appropriate to the given real-world situation, then the human agent who stipulated the robo-agent as computing moral view B has an excuse, since the robo-agent malfunctioned and computed moral view A.

If a human agent suffers loss and demands satisfaction in a court of law, the human agent who stipulated the robo-agent computed moral view A (and who bears liability responsible for the behavior of the robo-agent) might lie and say that he/she had intended to stipulate the robo-agent as computing moral view B, but was mistaken as to what was moral view A and moral view B, conflating them. What evidence could be used to decide that the human agent is lying or telling the truth? To prevent scenarios such as this, it might be required that any human agent who interacts computationally and socially with robo-agents sign an agreement before the interactions occur that he/she stipulates what the robo-agent does, and that he/she understands the difference between the different stipulations that might be made (such as the stipulation of moral view A and the stipulation of moral view B). Of course, to make such an agreement, the human agent must understand moral view A and moral view B, the program implementing moral view A, and how a normal condition in computing moral view A is at the same time a breakdown condition in computing moral view B. In short, each human agent needs to understand Kripke's argument.

Thus, to prevent the courtroom scenario where a human agent lies about making a stipulation, that human agent must understand Kripke's argument in order to sign (and make legal) the agreement. But where human agents do not understand Kripke's argument, they fail to satisfy the criterion of knowledge for liability responsibility. If so, they lose the capacity for liability responsibility. The dilemma is that either there are no safeguards against lying in court (in the kinds of situations described above) or there are safeguards, but under them most stipulators lose liability responsibility—and thus would not need to appear in court in such situations (such as the one described above).

There are additional worries about liability responsibility. Is stipulating what it is that a robo-agent does a criterion—perhaps a necessary and sufficient condition—for the intentions of a human agent? Similarly, can the stipulations of a human agent show whether or not the action of a robo-agent is an accident? If a human agent has to stipulate the behavior of a robo-agent as being such-and-such because the corporation that designed the robo-agent requires such a stipulation, is the human agent closed off from making a choice?

6.2. Capacity Responsibility

If we decide on the basis of agreed-upon criteria (which could be extraordinarily complex and difficult for many human agents to understand) whose stipulation counts, do we want to say that some human user is responsible for what it is that a given robo-agent does? For one, the criteria might be so difficult to understand that almost any human agent who satisfies those criteria might not understand them [25]. If so, the human agent fails to have capacity responsibility for the actions of the robo-agent. Similarly, if the human agent does not understand the program the robo-agent implements, can that human agent understand the stipulation that is made as to what it is that the robo-agent does? Finally, must a human agent who makes a stipulation as to what it is a robo-agent does understand Kripke's argument? If not, it *prima facie* appears the human agent does not understand why such a stipulation has to be made. If so, how should it be determined that a human agent understands Kripke's argument?

Clearly, the standards of understanding what one does in making a stipulation must be high. But might they be so high that many human agents will fail to attain them? Would a court of law decide that they should be no higher than what a reasonable person could understand? What would that mean? What does a reasonable person understand? If Kripke's argument is sound, there will be many difficult problems attending the criteria that are employed for ascribing capacity responsibility to human agents who computationally and socially interact with robo-agents.

6.3. Causal Responsibility

Establishing causal connections between agents and events can depend upon how the events are individuated. Suppose that human user A asserts that a given robo-agent has successfully performed moral action A, while human user B asserts that the same robo-agent has unsuccessfully performed moral action B. What action has the robo-agent performed? If it turns out that the human agent who makes the stipulation as to what a robo-agent does is the cause of the robo-agent doing that, how would it be decided that it is the word of human user A that counts, while the word of human user B does not count? What would be the conditions necessary for declaring the word of so-and-so to count in deciding what it is that a given robo-agent does? Could those conditions be contested in a court of law? Imagine a specific scenario: robo-agent A is stipulated as successfully initiating a shut-down of a city subway system after it is discovered that a hostile, armed militia intent upon terrorizing that city is riding a subway car in a certain section of the city. However, robo-agent A could also be stipulated as unsuccessfully blocking any attempt to shut down the city subway system. In deciding what it is that the robo-agent did, whose voice counts? Here the general problem is that of determining causal responsibility. Typically, non-human agents (such as robo-agents) can be the cause of an event. However, can robo-agents be causally responsible for an action if what it is they do is the result of a stipulation made by a human agent? Can causal responsibility be partially transferred from a human agent (who stipulates what it is that a robo-agent does) to a robo-agent? What does it mean to say 'partially transferred?' How are robo-agents individuated? These are just some of the questions about causal responsibility that must be addressed if Kripke's argument is sound.

6.4. Human Agent Responsibility for All the Actions of a Robo-Agent

Since some human agent A who uses robo-agent software (not necessarily the designer of the software) will need to stipulate all of the actions that a robo-agent performs, A might be responsible for all of the actions of a given robo-agent, depending upon how the issues regarding liability, capacity, and causal responsibility are resolved. Suppose that A purchases a software package for a robo-agent in order that the robo-agent performs certain actions, but not all of those actions that the robo-agent could perform given that software poackage. A has an interest in the robo-agent performing certain actions, but no interest in the robo-agent performing certain other actions. If the robo-agent performs those certain other actions (because someone else is using the software—who has not made any stipulations about the actions of the robo-agent) and the result is an injury sustained by some human being, is A responsible for that injury? If stipulations are made for only a subset of actions of a robo-agent, the human user who makes the stipulations could stipulate only those actions in which he/she has an interest. But since *any* action of a robo-agent needs to be stipulated by some human agent, and since some of the actions a robo-agent can perform can depend upon other actions that they perform (often in novel and unanticipated ways), it appears that stipulations must not be restricted to a subset of actions of the robo-agent. Rather, the full set of actions the robo-agent could perform (whether it does or does not actually perform them) should be the proper object of stipulation.

These considerations raise several questions. Should a human agent (whose voice does not count in making stipulations as to what a robo-agent does) be partially responsible for some of the actions of a robo-agent with which the human agent computationally and socially interacts? If a stipulation as to what it is that a robo-agent does requires the agreement of an entire community, then must each member of that community bear some responsibility for some subset of the actions of the robo-agents with which they computationally and socially interact (there are difficult philosophical problems concerning community-wide agreement [25])? Or should human users of robo-software whose voices count in determining what it is that a robo-agent does be solely responsible for that subset of actions that robo-agents perform?

7. Does Kripke's Argument Have Practical Import for the Field of Robo-Ethics?

It is important to anticipate one kind of objection that could be made by non-philosophers (especially those doing research and development in robo-ethics) who might criticize the kinds of concerns I raise as being merely philosophical, and thus having no practical impact for the field of robo-ethics.

7.1. Wallach and Allen

For example, critics such as Wallach and Allen [26] (who are philosophers) have suggested that we can avoid worrying about some of the philosophical concerns that philosophers raise with respect to questions about, say, moral agency and autonomy in the context of robo-ethics (or what they call "machine ethics"). Instead, Wallach and Allen argue that all we need to do is to develop moral machines capable of being good moral reasoners. These robo-agents exhibit what Wallach and Allen call 'functional morality', [26] in contradistinction to robo-agents that exhibit 'operational morality.' Simply put, the distinction is between robo-agents that are themselves capable of sophisticated moral reasoning ('functional morality') and robo-agents whose (low-level) moral reasoning is hard-wired in by the software designer ('operational morality'). Gunkel makes the point neatly: " ... while we busy ourselves with philosophical speculation concerning the moral status of the machine, machines are already making decisions that might have devastating effects for us ... rather than quibbling about obscure metaphysical details or epistemological limitations ... we should work with and address the things to which we do have access and can control [27]". But, as we have seen in the arguments above, in developing moral machines (i.e., robo-agents) two new philosophical problems arise for both the design of software for robo-agents and for the implementation of robo-agents in physical computers. These two philosophical problems affect robo-agents that exhibit functional morality, and those that exhibit operational morality and they cannot be ignored by anyone who is involved in robo-ethics—research and development, marketing, sales, local and national policy matters, users (corporations, small businesses, individuals). One cannot rationally adopt a policy of 'don't care' about these two problems.

Tavani [28] (p. 6), in his critical review of *Moral Machines* [26]), notes that Wallach and Allen define robo-ethics—or what the authors call 'machine ethics'—as a "field that extends or expands upon computer ethics in at least two ways: (1) by shifting the concern away from 'what people do with computers to questions about what machines *do by themselves*', and (2) by 'fostering a discussion of the technological *issues involved in making computers themselves into explicit moral reasoners*'" [italics mine]. But the arguments in this paper show that any discussion about what robo-agents do by themselves is moot—what they do is determined by what human agents stipulate that they do. So if Kripke's argument is sound, the first prong in Wallach and Allen's definition of robo-ethics is no longer tenable. Also, since understanding the technological issues involved in both simulating human moral reasoning and robo-agents engaging in moral reasoning (which might not be a simulation of human moral reasoning) requires understanding and resolving the two new philosophical problems for robo-ethics, the second prong in Wallach and Allen's definition of robo-ethics requires understanding and resolving them.

7.2. Anderson and Anderson

Anderson and Anderson [29,30] have developed code for moral reasoning based on Ross's deontological framework of "prima facie duties" (for an extended discussion/analysis of functional morality as it relates to frameworks found in Anderson and Anderson [29,30], as well as in Wallach and Allen [26], see Tavani [31]). Moral agents implementing such code might exhibit functional morality. No matter whether they exhibit functional morality or operational morality, however, Kripke's argument applies to them—both the physical implementations of such robo-agents and the design of the moral reasoning systems depend upon resolving the two new philosophical problems for robo-ethics. Moreover, testing procedures for such software are themselves programs subject to the same difficulties. Both verifying that the program is correct by inspecting and reasoning about the program and testing the program are targeted by Kripke's argument.

S. Anderson [32] (p. 30) has argued that the ultimate goal of robo-ethics is to build a system that "follows an ideal ethical principle or set of principles in guiding its behavior". She then describes various levels at which machines can be designed to act ethically. Tavani [31] (p. 337) has suggested that Anderson can be interpreted as identifying three distinct software design levels at which robo-agents behave ethically: (i) limit the moral behavior of the robo-agent to prevent it from causing moral harm; (ii) provide instructions for behaving morally in specific ways; and (iii) provide robo-agents with ideal moral principles as well as learning procedures so that they can learn how to apply the principles in specific situations. No matter which of these design stages one considers, however, each is targeted by Kripke's argument, as are their physical implementations in physical computers.

7.3. The Moral Turing Test (MTT)

Testing the moral behavior of a robo-agent (regardless of its design level) is problematic when using what Allen et al. [33] call a "Moral Turing Test" (or MTT) (as implied, MTT extends the classic Turing test to include questions pertaining to moral reasoning). But MTT is also targeted by Kripke's argument. Consider that the responses of the robo-agent to the questions constituting the MTT are programmed responses—that is, some computer program makes the responses and that program (which must be physically implemented in a physical computer) constitutes the robo-agent making such responses. If Kripke's argument is sound, MTT will fail to provide an adequate test for modeling human-like moral reasoning in robo-agents. This would put the framework of functional morality under considerable pressure, since there would be no means of testing the moral capabilities of robo-agents that *excludes* the stipulation of a human agent that the robo-agents do have such-and-such moral capabilities. Robo-agents are not being tested by MTT for such-and-such moral capabilities—they are being declared outright by stipulation to have them.

7.4. Alternative Views of the Field of Robo-Ethics

Veruggio and Abney [34] argue that there are three different ways of defining robo-ethics. This paper considers all three definitions, since someone might object that on one or another of the three definitions of robo-ethics, Kripke's argument does not raise any new philosophical problems for robo-ethics, and it is that definition of robo-ethics that either is used or should be used by people working in the field. Their first definition is that it "applies to the philosophical studies ... about the ethical issues arising from the effects of the application robotics products on our society [34]". Their second definition is that robo-ethics is "the moral code to which the robots themselves are supposed to adhere [34]". The third definition is that robo-ethics refers to "the self-conscious ability of the robots themselves to do ethical reasoning ... and to freely, self-consciously choose their course of action [34]". Regarding the first definition, the two new philosophical problems for robo-ethics arise for both the ethical issues and the philosophical studies of those issues—see, in particular, Sections 5.1–5.3 above. Regarding the second and third definitions, the two new philosophical problems for robo-ethics arise in both the design and the physical implementation of robo-software. Thus, on all three definitions of robo-ethics the two new philosophical problems for robo-ethics arise, and so must be resolved (if Kripke's argument is sound).

8. Conclusions

This paper has argued that Kripke's argument raises two new philosophical problems for robo-ethics and that these two philosophical problems raise difficulties throughout the field of robo-ethics. In particular, (i) a definition of robo-ethics (or "machine ethics") proposed by Wallach and Allen [26] needs to be reformulated; (ii) we know what actions a robo-agent performs only because some human agent stipulates that it performs those actions; (iii) many ethical questions in robo-ethics need to be reformulated; (iv) some philosophical arguments for philosophical positions within the field of robo-ethics need to be reformulated; and (v) the concepts of legal and moral responsibility used in robo-ethics need to be reformulated. If Kripke's argument is sound, then researchers in the field of robo-ethics need to respond accordingly.

Funding: This research received no external funding.

Acknowledgments: I would like to thank Herman Tavani for reading several versions of this manuscript and for making extensive and important helpful comments and suggestions and an anonymous referee for motivating me to make the arguments of this paper more transparent by asking important questions about them (in an earlier version).

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

References

1. Levin, J. *Functionalism*; Stanford Encyclopedia of Philosophy: Palo Alto, CA, USA, 2018.
2. Avigad, J.; Blanchette, J.; Klein, G. Introduction to *Milestones in Interactive Theorem Proving*. *J. Autom. Reason.* **2018**, *61*, 1–8. [CrossRef]
3. Sitaraman, M. Building a push-button RESOLVE Verifier: Progress and Challenges. *Form. Asp. Comput.* **2011**, *23*, 607–626. [CrossRef]
4. Avigad, J. Formally Verified Mathematics. *Commun. ACM* **2014**, *57*, 66–75. [CrossRef]
5. Buechner, J. Not Even Computing Machines Can Follow Rules: Kripke's Critique of Functionalism. In *Saul Kripke*; Berger, A., Ed.; Cambridge University Press: New York, NY, USA, 2011.
6. Buechner, J. Does Kripke's Argument Against Functionalism Undermine the Standard View of What Computers Are? *Minds Mach.* **2018**, *28*, 491–513. [CrossRef]
7. Buechner, J. *Gödel, Putnam, and Functionalism*; MIT Press: Cambridge, MA, USA, 2008.
8. Putnam, H. *Reason, Truth, and History*; Cambridge University Press: New York, NY, USA, 1981.
9. Ryle, G. *The Concept of Mind*; University of Chicago Press: Chicago, IL, USA, 1983.
10. Kaufman, S.; Rosset, S.; Perlich, C. Leakage in Data Mining: Formulation, Detection, and Avoidance. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'11, San Diego, CA, USA, 21–24 August 2011.
11. Christiano, P.; Leike, J.; Brown, T.; Martic, M.; Shane, L.; Amodei, D. Deep Reinforcement Learning from Human Preferences. *arXiv*, 2017; arXiv:1706.03741v3.
12. Popov, I.; Heess, N.; Lillicrap, T.; Hafner, R.; Barth-Maron, G.; Vecerik, M.; Lampe, T.; Tassa, Y.; Erez, T.; Riedmiller, M. Data-Efficient Deep Reinforcement Learning for Dexterous Manipulation. *arXiv* **2017**, arXiv:1704.03073.
13. McCarty, L.T. Intelligent Legal Information Systems: Problems and Prospects. *Rutgers Comput. Technol. Law J.* **1983**, *9*, 265–294.
14. Shortliffe, E. *Computer-Based Medical Consultations: MYCIN*; Elsevier: Amsterdam, The Netherlands, 1976.
15. Weis, S.; Kulikowski, C.; Amarel, S.; Safir, A. A Model-Based Method for Computer-Aided Medical Decision-Making. *Artif. Intell.* **1978**, *11*, 145–172. [CrossRef]
16. McCarty, L.T. Reflections on TAXMAN: An Experiment in Artificial Intelligence and Legal Reasoning. *Harv. Law Rev.* **1977**, *90*, 837–893. [CrossRef]
17. Anderson, M. GenEth. Available online: http://uhaweb.hartford.edu/anderson/Site/GenEth.html (accessed on 12 June 2018).
18. Tavani, H.T. Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. *Information* **2018**, *9*, 73. [CrossRef]
19. Jonas, H. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*; University of Chicago Press: Chicago, IL, USA, 1984.
20. Coeckelbergh, M. Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* **2010**, *12*, 209–221. [CrossRef]
21. Gunkel, D.J. The Other Question: Can and Should Robots Have Rights? *Ethics Inf. Technol.* **2017**, *19*, 1–13. [CrossRef]
22. Moor, J.H. Four Kinds of Ethical Robots. *Philos. Now* **2009**, *17*, 12–14.
23. Audi, R. (Ed.) *The Cambridge Dictionary of Philosophy*, 2nd ed.; Cambridge University Press: New York, NY, USA, 1999.
24. Coleman, J. *Risks and Wrongs*; Cambridge University Press: New York, NY, USA, 1992.
25. Kripke, S. *Wittgenstein On Rules and Private Language*; Harvard University Press: Cambridge, MA, USA, 1982.

26. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Right from Wrong*; Oxford University Press: New York, NY, USA, 2009.
27. Gunkel, J. *The Machine Question*; MIT Press: Cambridge, MA, USA, 2012; p. 75.
28. Tavani, H.T. Can We Develop Artificial Agents Capable of Making Good Moral Decisions? *Minds Mach.* **2011**, *21*, 465–474. [CrossRef]
29. Anderson, M.; Anderson, S.L. A Prima Facie Duty Approach to Machine Ethics. In *Machine Ethics*; Anderson, M., Anderson, S.L., Eds.; Cambridge University Press: New York, NY, USA, 2011; pp. 476–494.
30. Anderson, M.; Anderson, S.L. Case-Supported Principle-Based Behavior Paradigm. In *A Construction Manual for Robot's Ethical Systems*; Trappl, R., Ed.; Springer: New York, NY, USA, 2015; pp. 155–168.
31. Tavani, H.T. *Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing*, 5th ed.; John Wiley and Sons: Hoboken, NJ, USA, 2016.
32. Anderson, S.L. Machine Metaethics. In *Machine Ethics*; Anderson, M., Anderson, S.L., Eds.; Cambridge University Press: New York, NY, USA, 2011; pp. 21–27.
33. Allen, C.; Varner, G.; Zinser, J. Prolegomena to Any Future Moral Agent. *Exp. Theor. Artif. Intell.* **2000**, *12*, 251–261. [CrossRef]
34. Veruggio, G.; Abney, K. Roboethics: The Applied Ethics for a New Science. In *Robot Ethics: The Ethical and Social Implications of Robotics*; Lin, P., Abney, K., Bekey, G., Eds.; MIT Press: Cambridge, MA, USA, 2012; pp. 347–363.