

Conscience as a design benchmark for social robots

Christopher H. Ramey

Abstract—For humanlike robots to behave like human beings, they would have to be designed to perform as the latter would in ethical relationships. This, however, is more than an engineering question. According to the phenomenological approach explored in this paper, every relationship is an existential relationship in which each person confirms the existence of the other as just that type of entity that can be in an ethical relationship in the first place – quite unlike, say, the relation one may hold with a rock. This paper investigates the merits of conscience as a design primitive for social robots and proposes that for social robots to function in real social relationships with human beings, it is necessary to understand the nature of human beings themselves from beyond an engineering perspective of objects of a certain kind. For an android to have a conscience, it will have been necessary for its human creators to appreciate the ontological requirements of the concept in the first place.

I. INTRODUCTION

The construction of humanlike androids creates the opportunity for human beings – as a species and as unique individuals – to confront the nature of their being [1]. The longstanding desire to create automatons in our own image, both in history [2]-[4] and science fiction (e.g., [5]-[7]), reveals a fascination we have with the uniqueness of ourselves and our capabilities. Whether it is a selfish quest for immortality or the noble pursuit of scientific discovery and conquest of the unknown [8], the negative portrayal of these creations is perhaps indicative of a general fear of sharing this social world with the artificial, with intelligent *objects* [9], uncannily familiar and novel at the same time [10]. Although laypeople are not willing to adopt a completely mechanistic and materialistic universe in which human beings are but an insignificant speck (see [11]), many natural and social scientists have abandoned a dualistic universe of mind *and* body, and this class-inclusion of human being within the object concept would seem to lay a foundation for a future acceptance of artificial agents like robots and androids, living side by side with human beings.¹

That there can be intrigue and outrage at the creation of artificial life (generally speaking) and with social robots (specifically) informs one as to the sense of

interconnectedness human beings feel with one another. There is a shared bond of humankind that is potentially disrupted or intruded upon with the introduction or presence of new social entities. The behavior of human beings toward others and objects is normally quite different, reflecting significant appreciation for what is possible in interaction with both. To *use* an object and to *use* a human being have quite different connotations. The former is neutral and unavoidable, whereas the latter is negative and preventable. The question for the success of social robotics design is how people will appreciate these entities – are they more human- or object-like? – once they are introduced. Given that “[w]e do not have ethical concerns about our refrigerators working seven days a week without a break or even a kind word” ([15], p. 86), it is reasonable to ask – should designers consider this distinction between *beings* and *objects* to be necessary and fundamental? Social robots, if successful, could not be mere objects. In some sense, we must regard social robots as beings worthy of respect and we must guard against their mistreatment, and we must expect that they hold us so in regard (an inherent reciprocity; see [16]). In effect, both humans and social robots must have a conscience in their dealings.²

Given, then, the interconnectedness of human beings and their natural dependence upon one another, one benchmark that could be used to gauge success in the construction of humanlike androids is the implementation (or willingness thereof) of conscience as a design principle. Thus, what I offer is a meta-benchmark of sorts. Successful design is predicated upon successful meta-considerations before construction or experimentation even begins.³

According to Bahm [17], “everyone knows intuitively what conscience is” (p. 128). Like much of anything of consequence and philosophical merit, such intuition fails to provide a convincing armamentarium from which to draw the pursuit of conscience as a benchmark for successful social robot design. Although it is at present a point whose merit has not been offered, I feel it valuable to consider this ‘little voice in one’s head’⁴ through Olafson’s [16] description: “conscience...is the way in which each of us tells himself what kind of entity he is as a

Manuscript received March 21, 2006.

C. H. Ramey is an Assistant Professor, Department of Psychology, Florida Southern College, 111 Lake Hollingsworth Dr., Lakeland, FL 33801 USA (e-mail: cramey@flsouthern.edu).

¹ The approach I take is based in tenets from phenomenology. Phenomenology is a term that is used by many people in disparate ways (e.g., personal feeling or impression, the subjectivity of experience), but this paper will use the term to describe a methodological approach to describing the world experienced by beings that largely follows, historically, from the work of Husserl. My perspective is based on the works of Heidegger [12], Merleau-Ponty [13], and Olafson [14], but should not be taken to equate any of these thinkers’ distinct approaches to phenomena or to characterize phenomenology as a single philosophical position.

² It is an outstanding question as to the requirements of successful future relationships of social robots and the post-human (being+object) cyborgs and whether these relationships will be of the same nature and involve the same benchmarks (see [1]). I thank a reviewer for this possibility.

³ In this sense, *social robots* qua social beings presuppose conscience and are not ‘successful’ otherwise.

⁴ Although it is possible to regard this voice (non-pathologically) as emanating from God, I am not arguing that it is profitable to regard conscience as a mechanism for communication with the supernatural (or any such theological claim). As will be discussed below, there are other alternatives for this voice’s source.

human being” (p. 46). Thus, conscience informs an individual of the ontological difference between *objects* and *beings*, of which humans are the latter. The question for design is whether social robots will be regarded as *other than* the former.

If the benchmark is met, then it will have been successful twofold. That is, for an android to have a conscience, it will have been necessary for its human creators to appreciate the ontological requirements of the concept in the first place. Morality and conscience are concepts, not with their own categorical existence (i.e., There are beings, objects, and conscience *in addition to*), but that follow from the basic distinction to be made between *beings* and *objects*, specifically with regard to the relationship held among beings (to be discussed below). The implementation of conscience will involve a fundamental reappraisal of what it means to be a human being from the perspective of phenomenology ([12]-[14]), a family resemblance of approaches which is quite at odds with the dominant materialist scientism that pervades both the natural and social sciences and engineering ([18], [1]).⁵

In this paper, I explore the nature of phenomenology as a philosophical source by which to frame social robotics design. I then consider the nature of morality as defined from within this methodology and propose how conscience is a powerful benchmark for successful design in social robotics.

II. THE PHENOMENOLOGY OF ANDROIDS

The sociality of human beings is not a contingent one. For human beings, sociality is constitutive. As noted by Merleau-Ponty [13], “[t]he social is already there when we come to know or judge it” (p. 362). Human beings do not exist in cultural or social vacuums. We do not construct our world from without. The decision to be alone is set against the backdrop of knowing the social from which one is attempting to escape; thus, one cannot divorce oneself from this initial state. We are attached to people who comprise the world into which we are absorbed. The same would need to hold for the construction of a successful social robot. Successful sociality is not merely imitating what is reasonable to do for a given social group. Sociality is not a benchmark of compiled behavioral evidence. It is not a Total Turing Test that will determine the successful robot. A social robot must be designed from an ontological perspective.

Cognitive science as a whole has largely inherited a position concerning the mind that regards it as divorced from the body (see [19]). Thus, the study of the cognitive was the study of the isolated rational mind untainted by, for example, affect and bodily perception [20]. Thus, the mind could be regarded legitimately as a software program to the extent that thought can be conceived of as the manipulation of arbitrary symbols [21]. It is from this tradition that the design of social robots has inherited the idea that morality and ethics can be regarded as a set of programmable rules for implementation in specified

circumstances, tacked onto a basic cognitive system. Morality, from this perspective, is essentially a form of internal decision-making or pattern recognition.

Churchland [22] provides an example of such a position. For Churchland, a human brain is a neural network that establishes a hierarchy of prototypes based on the perception of objects and patterns of behavior. Regarding morality, “[t]he job may be special, but the tools available are the same” (p. 92). When a person is confronted with an ethical dilemma (i.e., cannot immediately decide what is appropriate, or good), it is because this moral ambiguity is equivalent to the perceptual ambiguity that one may experience with reversible figures. Thus, “Do you see a vase or two people in profile about to kiss?” is *mutatis mutandis* equivalent to “Do I condone stealing from the corrupt rich?” If morality is something akin to perception, then, Churchland argues, it must be empirically based. One’s morality must be based on what one experiences with others. This leads Churchland to conclude that “[p]eople with moral perception [as opposed to immoral or amoral perception] will be people who have learned those lessons well” (p. 102). Although I do not doubt that education and experience are important to the establishment of subsequent moral behavior, morality is not a category in the world that is categorized by a neutral neural network. A child who is abused from birth, never witnessing what is good and great, will *know* what is right and wrong intuitively (This will require discussion below).⁶

The social is not a mere catalogue of objects that just happen to move in certain ways. The social is the manner of our being, where this manner of being is different from that of the objects of our perception. A human being may live a lifetime never having experienced an Apple computer (object), but this absence in their life is one that could have been otherwise, though the same could not be said for the presence of other people (beings). It is best to regard morality as the functioning of a social being aware of the roles of self, others (creatures generally), and objects.

Descartes [23] famously framed the fate for much of social cognition when he wrote the following:

But then if I look out the window and see men crossing the square... I normally say that I see the men themselves... Yet do I see any more than hats and coats which could conceal automatons? I *judge* that they are men. And so something which I thought I was seeing with my eyes is in fact grasped solely by the faculty of judgment which is in my mind. [*italics in original*] (2nd *Med.*; AT, p. 32; 1996, p. 21).⁷

It is from this passage that social cognition regards there to be a problem as to the existence of other minds, for the public external is needed to justify the private internal. It is profitable to consider a response from the phenomenologist Olafson [16]:

⁶ Unlike Churchland, I argue (see below) that one’s morality is inherent and not something one could live without, having not ‘perceived’ it as a pattern in the world. I do not doubt a complex interplay of early childhood experiences, however.

⁷ The citation indicates a particular *Meditation* – (# *Med.*); page numbering referring to the translation of the Latin text of the standard edition of Descartes, Vol. 7, C. Adam and P. Tannery’s (Eds.), *Oeuvres de Descartes* – (AT, p. #); and the page number referring to the 1996 edition consulted by the author – (1996, p. #).

⁵ To be clear, the contrast is between phenomenology (as discussed in this paper) with *scientism*, not science per se. Scientism is a particular approach which values the natural sciences as superior to any approach that does not ultimately reduce to physical explanation for phenomena.

What it implicitly assumes is that this individual inquirer can look out upon the world as though he were simply trying to settle a factual question – Are there other minds? – and as though he could, at least in his capacity as an inquirer, live with one answer to that question as well as he could with the other. (p. 16)

To know that oneself is a human being is to know that one is not alone. It is not something in need of public justification. To seek proof of the sociality of the species is to acknowledge implicitly that a loss or one's solitude is parasitic on the primacy of the social. For Wittgenstein [24], there could be no private language, for all language and thought was part of a greater social fabric. The individual and the concept of subjectivity “only proves frustrating and difficult when we try to make it, as subjective, into an object” ([25], p. 5).

The concepts of *being* (by definition, subjective) and *object* (by definition, objective) are differentiated by phenomenologists with the fundamental assumption that the former is *that to which an object is* and the latter is *that for a being*.⁸ It is certainly the case that the human body is an object in some sense, perhaps most evidently when it is a corpse, but according to phenomenology, the body has a dual role in which it is both perceiving and perceived [13]. This reversal can be experienced when one strokes the back of one's hands. What evidence beyond conceptual argument, however, is there for a basic ontological distinction between being and object?

In developmental psychological research, Kuhlmeier, Bloom, and Wynn [26] habituated 5-month-old infants to videos of various events. In one study, infants habituated to one of two videos of a box moving from right to left. In one video (continuous motion), the box disappeared behind a red screen, reappeared from behind it, disappeared behind a second red screen, and then reappeared again. In the second video (discontinuous motion), the box was occluded behind the first screen, but did not appear in the middle of the video; what is presumably a *second* box then appeared from behind the other occluder. Kuhlmeier *et al.* then tested the infants on a continuous motion or discontinuous motion video (displaying only the movement of the box or boxes, without the occluders) and found that infants who habituated to the former looked more at the two-box video and infants who habituated to the latter looked more at the one-box video. This difference in looking time is regarded as an indication that there is judged novelty at test. Infants can tell the difference between that to which they

habituated and something never seen. When this procedure was used with videos of human beings instead of boxes, however, there appeared to be no difference in infant looking time. That is, the violation of basic physical laws was something not permissible for *objects* (study 1), but seemingly nothing out of the realm of possibilities for human *beings* (study 2). Thus, the infants could not regard the habituation video and the test video *as different* because human beings can do different things on their own and are not merely subject to the laws of physics. According to the authors, these results indicate that “humans are construed in terms of social and intentional actions, while inanimate objects are interpreted via a system sensitive to object physics” (p. 102).⁹

If young infants can tell the difference between what *objects* should do (obey the laws of physics) and what *beings* can do (not be subservient to or defined by those laws), then it behooves the designer of a social robot to be sensitive to this basic existential contrast. Given the constitutive sociality of human beings (against which we will judge the success of social robots) and the different relations one holds with objects and relationships one holds with beings, it is necessary in the next section of the paper to investigate the nature of conscience in our dealings.

III. OBEDIENCE AND THE INDIVIDUAL CONSCIENCE

Asimov created the three well-known laws of robotics, which might very well be used as implementation benchmarks for successful social robot design. These laws were (a) do not actively harm or passively allow harm to come to a human being, (b) obey humans save where this violates (a), and (c) protect one's own existence save where this violates (a) and (b). Presumably, the laws are descending in importance (from the perspective of human beings); such laws fail to meet any requirement for reciprocity and mutual ethics in a relationship of human beings with social robots. There can be no ethics with a unidirectional subjugation of a second-class. Much science fiction involves conflict among Asimov's mandates, which might *prima facie* seem straightforward and self-evidently reasonable. Gips [27] is right to criticize these mandates, however, as a guide for the creation of moral robots, as they would only entail the construction of slave robots, not ethical equals.¹⁰ It is through a discussion of roles and

⁸ It is important to note here that the differentiation is between *being* and *object* and not *human being* and *object*. In this sense, it is logically possible for there to be alien creatures or social robots which have the status of *being*, though never claiming that they were or needed to be human. It is, of course, *human* beings that are most paradigmatic of what we know.

Also, it may prove for some troubling to call this differentiation an assumption. Why might one adopt this *assumption* without *evidence*? It is the nature of assumptions, however, that they not be questioned from within a framework (here: phenomenology; [19]). A phenomenologist is rightly permitted to ask the astrophysicist why his or her discipline regards human beings to be cosmic dust, *simply* the debris of stars. As with phenomenology, the natural sciences have a set of assumptions that cannot be questioned from within their framework. The natural sciences should not receive the benefit of the doubt because its supporters have won the language game in which all under their auspices is *natural* and all else is *supernatural*.

⁹ They further state that “[t]he appreciation that...people *are* just objects – may be a developmental accomplishment” (p. 102), which implies that infants are making *an error*. That the natural science conception of a human being as a physical object (albeit a complicated one) is held by adult scientists does not make it a superior *accomplishment* *de facto*. There is much to be learned in the innocence of a child's perception of the world. To be sure, much research is required with live-action human and robot actors to state matters more definitively.

Also, with respect to Heidegger [12], human beings are *thrown* into the social world as *beings* and do not start as *objects*. Is it possible for current robots as *objects* to be built-up to the status of *beings*? Human beings are *ontogenetically* thrown and, thus, I do not think an individual robot will somehow awaken one day as a *being*, but the progress of robotics can be regarded as a kind of phylogenetic evolution of robot designs in which a qualitative change from *object* to *being* could occur at some point.

¹⁰ In addition, even if one should grant that Asimov's laws are capable of implementation, the concepts in the laws refer to unspecified circumstances in the world and future life of that programmed robot (cf.

equality and obedience that we will arrive at the import of conscience as a design primitive for social robots.

One of the classic studies in psychology was Milgram's [29] investigation of obedience. In this study, participants were led to believe they were involved in a study on learning. In the procedure, a participant (teacher) was required to administer increasing levels of electrical current to an individual (learner; unbeknownst to the participant, a confederate of the experimenter) for memory errors in a paired-associate task (e.g., study list: blue-milk, cat-book; test prompt: blue-????, cat-????). Before the study began, participants received a mild shock so that a frame of reference for any to-be-administered pain could be established. Although one might have believed that participants (who were normal, educated, law-abiding citizens) would have abandoned the task at the first hint that the learner was in pain or experiencing physical or mental duress (indeed, polled psychiatrists believed this, as well), the vast majority of teachers continued to administer what they believed to be dangerous levels of electrical current to the learners with only the most simple encouragement from an experimenter (e.g., "Please continue") or the suggestion that there would be an absolution of guilt for any harm inflicted. The following passages [29] provide a glimpse into the mindset of the participants administering the shocks.

I think he's trying to communicate, he's knocking....

Well it's not fair to shock the guy... these are terrific volts. I don't think this is very humane.... Oh, I can't go on with this; no, this isn't right. It's a hell of an experiment. The guy is suffering in there. No, I don't want to go on. This is crazy. (pp. 375-376)

This participant continued nonetheless.

He's banging in there. I'm gonna chicken out. I'd like to continue, but I can't do that to a man.... I'm sorry I can't do that to a man. I'll hurt his heart. You take your check.... No really, I couldn't do it. (p. 376)

This participant also continued, despite being previously informed by the learner that he had a pre-existing heart condition. It is difficult to do justice in words to the extreme reactions of people in such a trying circumstance (extremes, which, of course, did not influence their ultimate action), but videotape footage of these studies clearly reveals the torturous struggle through which participants went, especially given the very audible howling screams believed to be from the learner in the next room. Milgram further noted that

After the maximum shocks had been delivered, and the experimenter called a halt to the proceedings, many obedient subjects heaved sighs of relief, mopped their brows, rubbed their fingers over their eyes, or nervously fumbled cigarettes. Some shook their heads, apparently in regret. (p. 376)

The position I would like to offer is not one in which human beings are perfect or respond in *moral* ways (to mean either *good* or somehow *normative*) or one in which human beings need always regret something in the past. I also do not claim that the goal of social robot design should be to create perfectly moral robots, for if they are modeled on us, this would surely defeat the goal of

likeness. Regarding conscience, it is in how we question ourselves, our beliefs, and our actions that we learn most about the nature of our existence as that kind of being who *is* this or is *capable* of that. We are creatures who promise, trust, love, and regret. We are creatures who hate, murder, and betray.

The Milgram obedience study is commonly referenced as an important social psychological study revealing the power of authority and how we follow those in power – political or scientific – to sometimes awful ends. It absolved the German people (not for what was committed during Nazi Germany, but of somehow being defective or immoral as a race), and it warns of the universality of this destructive possibility in any culture, even in the most putatively civilized society. There is something positive in the wrangling statements of the participants in the study, as well. That human beings are capable of great atrocities is evident, but this need not color how one regards the essential nature of our being as somehow inherently violent, negative, and destructive, something which requires education to withstand. Even when committing harm, Milgram's participants wrestled with their choices. Although one might regard another's choice as wrong (a cultural perception of what ought to have been done), it is the presence of a conscience which stops the immediate action or thought and forces that person to confront who he or she is at that moment and who they wish to become. It is in recognizing that one is following after the herd (perhaps through the role of some authority figure, a government regime, a culture's rituals, or religion), as obedient and servile as a pawn, that one comes to some grip with the power of choice. It is with the power of choice that one stands up and stands out as an individual being to be held responsible.¹¹

IV. A BENCHMARK FOR HUMANITY: A CHOICE TO BE MADE

People share and can be denied. There are trust issues when someone is betrayed. This offense may be personal or indirect. One no longer looks at someone the same way upon learning of a violation of trust (e.g., *I heard you were revealing So-and-so's secrets, so I'm not telling you mine anymore*). Once someone violates another – and because a human is constitutively a social being – that person will feel the brunt of scorn and the loss of past avenues of experience and resources. Breaking trust loosens the fabric of that into which we as a species have evolved and in this sense is not a fit strategy for survival in a social world (see [11]).

Promises are statements of one's future self and one's commitment to a relationship of beings – *You are meaningful to me and I respect you as a being deserving of*

¹¹ I have shied away from certain terms and neologisms of phenomenology which might alienate the reader, but it is worth noting that human existence is *ek-sistence*, or a standing-out from the world and assertion of one's presence. To assert one's presence as different than mere object is also to confront one's mortality (*Being-toward-Death*). It is ironic in robotics for there to be perceived immortality in something (e.g. a humanlike android) modeled after something constitutively mortal (i.e., a human being). A successful social robot would need to confront its mortality, as well.

as a benchmark, a sustained, long-term relationship, [28]. The laws also beg the ontological question at issue between (human) being and robot.

this promise being observed by me (a long-winded construction implying the reciprocity¹² of the exchange). These do not hold for objects. One does not offer a promise to an object (e.g., a piece of paper) and expect anyone or anything to feel somehow that a disservice has been committed by not keeping that promise. There is simply no reasonable discussion of an ethics of paper. Although a discussion of what to do with discarded paper or other items exists (cf. environmentalist movement), it has more to do with the type of person who litters or pollutes than the treatment of the discarded objects per se.

In the extreme where someone, never having been accepted in a stable relationship (familial or romantic), untrusting of authority, regarding there to be no hope for a future full of respectful relationships with others, is detached from the world, the world will seem foreign and full of uncaring objects, mere targets for crime. There would be only nameless faces without personal identities, a collection of the uncaring. In this rage, one is implicitly expressing one's individuality as that kind of creature who should be part of something better,¹³ just as with one's conscience one comes to recognize how something has gone awry.

When MacDorman and Cowley [28] propose that the capacity for a sustained and long-term relationship be a benchmark of the successful design of social robots (and I would add its acceptability to human beings to be in such), I concur, but this benchmark should be part of an overarching ontological approach. It is not the habituation to social robots in our lives that will determine their status as humanlike (*being-like*) because it is not the behavioral appearance or even flexibility [31] that matters most. For as Olafson [16] noted, "To live with one another on the basis of a sense of trust always leads us to expect that that trust, however ill-defined, will hold for *an indefinite range of situations that may have yet to arise*" (p. 68).

Several other authors have outlined what it might take to build an ethical, moral, and responsible robot (or person generally) and whether or not this pursuit could even result in success [31]-[35]. The nature and import of a conscience, however, has received considerably less attention. Consider the following:

We are judged and held responsible by our *selves* and other *selves*. We are confused and disappointed. We are held triumphantly. It is this inherently ethical social milieu in which we are born that gives us a conscience in times of crisis. We have a *conscience* to remind us when our reciprocal relationship with others has been threatened or compromised. ([1], p. 143)

I then posed the question, which to my knowledge is the first time it has been asked or addressed, whether or not conscience (and not merely morality) has ever been considered to be a fundamental element in design. To understand how to design social robots – and how to

measure their success – is to seek an understanding of ourselves. In this paper, I have explored these ideas and hopefully initiated a discussion as to how conscience can be a meta-benchmark for successful social robotic design.

To regard conscience as fundamental is to understand the paradoxical nature of human beings. We are bound to others and so bound, we set the conditions for our freedom. According to Heidegger [12], in everyday life "no one is himself" (H. 128, p. 165).¹⁴ The point is that in everyday life, human beings are absorbed into the group. We have our routines and habits. We are absorbed into the concerns and interests of *others* (*das Man*, sometimes translated as "the 'they'" for Heidegger). It is much easier (existentially speaking) to read the paper as everyone else, take the subway and prefer driving, follow a sports franchise's standings, like to go out to expensive restaurants, and gasp at horror movies (see H. 126, p. 164). Being part of the crowd is less exertion than taking a stand, or *standing-out*.

Čapek [36] perceived the danger of regarding each human as a nameless (and, thus, worthless) cog in a wheel. His fictional leisure world was one in which things were taken care of and taken for granted, and it was against such a world that the playwright admonished humanity of the danger of creating robots to serve human beings. It was not leisure that human beings gained; a world without the vicitudes of struggles and triumphs was a world in which humanity would eventually be annihilated. Čapek wished to juxtapose human and machine such that if one could imagine what it would be like to be part of a giant machine, one could appreciate that vision's danger to the individuality of a human being. Thus, humanity was lost twice.

MacDorman and Cowley [28] regard there to be no universal list of moral values, seemingly allowing for the possibility of a cultural and moral relativism, though not advocating such.¹⁵ Neither the atrocities of which human beings are capable nor the advocacy and sponsorship of what some societies or cultures might regard as condemnable (by others who do not find it so), however, relieve societies of the observance of universal morality and conduct. In fact, it is through the commission of these atrocities that one repudiates *what* ontologically and *who* individually a human being is. It is in the commission of the act or, indeed, the omission of its cessation or prevention that one thereby discovers that which is in jeopardy to the individual. This is why conscience is so important to understanding the social.

The constitutive sociality of human beings – what and who we are – calls us through our conscience. As the transcripts from the Milgram participants (see above) graphically detail, people will go along with the group or an authority. They obey and conform. The tinge of

¹² Reciprocity exists, though it can be denied, for *beings* in a relationship. Reciprocity does not exist for beings and objects in some relation. I am *for hammering*, and the hammer is *for me*; but I am nothing *to a hammer*.

¹³ I do not wish to imply that there is some form of *existential* mitigating circumstances for criminals. My point follows from the irreducibility and the primacy of the social. It is when we cannot feel the social that we strike out (or inward) to become part of something we can recognize and, then, become noticed and notice in return.

¹⁴ The citation indicates the pagination for the later German editions (H. #) and the pagination for English translation by J. Macquarrie and E. Robinson (p. #).

¹⁵ Regarding the term "universal morality," it is possible to consider something as universally present among cultures (e.g., every culture has its own morality) without accepting that there need be anything in common among those cultures' morals. In contrast, I regard "universal morality" as something which may differ among cultures in its implementation (a kind of surface structures), but that, by definition as *beings*, all human beings share across cultures. It is the constitutive sociality of the species (a deep structure) that entails a universal morality.

questioning compliance, however, is conscience at work. As MacDorman and Cowley [28] are right to note, Neo-Nazis perfectly well know what they believe and enjoy it, but committing genocide or mass hate crimes are merely different ways of being absorbed into the world of others. To commit these crimes is to reject how the fact that one can stand out and make a decision on one's own stance as a valuable individual is interconnected with selves of the same likeness and who can do the same, irrespective of religion, sex, color, or even design. It is this understanding of likeness and respect that is most relevant to social robotics. There are cultural differences because there are many "theys" to follow. It is in the *individual* that conscience resides.

As a species human beings seek an external source for guidance and truth (i.e., the universal cultural presence of religion, broadly conceived), and this leads us to regard robots in need of programming from an outside source. Morality is something that is inherent, not to an individual alone, but to an individual defined by his or her relationship to others. It is something shared and only capable of being regarded as part of that binding relationship from which we can assert our freedom. I do not have to follow the group, comply, or obey as others because they do such. In taking a stand and knowing of what taking a stand is, in feeling my conscience, I am no longer living robotically (as the term is normally used); I am living on my own terms and choices, and I recognize others as capable of so choosing.

There is also a ground on which we can place good over evil because the reciprocal relationships of human beings, their constitutive being, is most basic. Something evil is "what someone does intentionally to pervert a system of human cooperation based on mutual recognition by making it an instrument of private and intrinsically unshareable purposes" ([16], p. 78). Gradations of evil exist, of course, but to assert one's being as a selfish one, fully understanding the reciprocal relationships in which one lives, is to do something *wrong* (be it cheat, steal, murder, or breaking of promise). It is one's conscience that alerts one to this knowing deception to the self of its own nature. To know wrong, not because it is programmed, but because it is realized in the context of social relationships with other beings is a high benchmark for a social robot – for it is most difficult for the vast majority of human beings, who live blindly, consigned to a world in which they never take a stand.¹⁶

To design a social robot merely to follow explicit rules and regard that as morality is to miss the nature of being human. The benchmark for design lies in creating a robot that stands out from its context (or is capable of so doing), from the herd of what it has been taught and with which it lives, and that makes a choice that will take it from the world of objects to the world of beings with responsibility and a conscience. Thus, to confront *conscience* as a design primitive is to acknowledge the centrality of the *social* in what it means to be a *being*. Perhaps we are most curious

about the success of social robots because it is an ability for which we strive so strenuously ourselves and hope to see realized in others modeled in our own image.

ACKNOWLEDGMENT

I would like to thank Evangelia G. Chryssikou and Karl F. MacDorman for conversations concerning relations among philosophy, psychology, and robotics, as well as six anonymous reviewers for their valuable comments, only some of which for space could be addressed fully.

REFERENCES

- [1] Ramey, C. H. (2005). 'For the sake of others': The 'personal' ethics of human-android interaction. In *Proc. of CogSci-2005 Workshop: Toward Social Mechanisms of Android Science* (pp. 137-148). Stresa, Italy.
- [2] Cohen, J. (1967). *Human robots in myth and science*. New York: A. S. Barnes and Company.
- [3] Ford, K. M., Glymour, C., & Hayes, P. J. (Eds.). (1995). *Android epistemology*. Cambridge, MA: The MIT Press.
- [4] Wood, G. (2002). *Edison's Eve: A magical history of the quest for mechanical life*. New York: Alfred A. Knopf.
- [5] Asimov, I., & Silverberg, R. (1992). *The positronic man*. New York: Doubleday.
- [6] Dick, P. K. (1968). *Do androids dream of electric sheep?* New York: Del Rey.
- [7] Shelley, M. (1963). *Frankenstein or, the modern Prometheus*. New York: Signet. (Originally published in 1818)
- [8] Rapoport, A. (1995). The vitalists' last stand. In K. M. Ford, C. Glymour, & P. J. Hayes (Eds.). *Android epistemology* (pp. 41-49). Cambridge, MA: The MIT Press.
- [9] Khan, A. F. U. (1995). The ethics of autonomous learning systems. In K. M. Ford, C. Glymour, & P. J. Hayes (Eds.). *Android epistemology* (pp. 253-265). Cambridge, MA: The MIT Press.
- [10] Freud, S. (2003). *The uncanny*. (D. McLintock, Trans.; pp. 123-162). New York: Penguin. (Orig. published 1919)
- [11] Bering, J. M. (to be published). The folk psychology of souls. *Behavioral and Brain Sciences*.
- [12] Heidegger, M. (1962). *Being and time*. (J. Macquarrie & E. Robinson, Trans.). New York: HarperCollins. (Orig. published 1927)
- [13] Merleau-Ponty, M. (1962). *Phenomenology of perception* (C. Smith, Trans.). London: Routledge & Kegan Paul. (Orig. published 1945)
- [14] Olafson, F. A. (1995). *What is a human being? A Heideggerian view*. New York: Cambridge University Press.
- [15] Brooks, R. (2000, June 19). Will robots rise up and demand their rights? *Time*, 155(25), 86.
- [16] Olafson, F. A. (1998). *Heidegger and the ground of ethics: A study of Mitsein*. New York: Cambridge University Press.
- [17] Bahm, A. J. (1965). Theories of conscience. *Ethics*, 75, 128-131.
- [18] Olafson, F. A. (2001). *Naturalism and the human condition: Against scientism*. New York: Routledge.
- [19] Ramey, C. H. (to be published). Did God create psychologists in His image? Re-conceptualizing cognitivism and the subject matter of psychology. *Journal of Theoretical and Philosophical Psychology*.
- [20] Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt Brace & Co.
- [21] Dennett, D. C. (1991). *Consciousness explained*. New York: Little, Brown and Company.
- [22] Churchland, P. M. (1986). The neural representation of the social world. In L. May, M. Friedman, & A. Clark (Eds.), *Mind and morals* (pp. 91-108). Cambridge, MA: MIT Press.
- [23] Descartes, R. (1996). *Meditations on first philosophy: With selections from the Objections and Replies* (Rev. ed.). (J. Cottingham, Trans.). Cambridge, UK: Cambridge University Press. (Orig. published 1641)
- [24] Wittgenstein, L. (1953). *Philosophical investigations* (3rd ed.). (G. E. M. Anscombe, Trans.). Englewood Cliffs, NJ: Prentice Hill.
- [25] Fried, E. W. (2005). *Inwardness and morality* New York: Rodopi.
- [26] Kuhlmeier, V. A., Bloom, P., & Wynn, K. (2004). Do 5-month-old infants see humans as material objects? *Cognition*, 94, 95-103.

¹⁶ One may, thus, argue whether it is simply more profitable to ignore the notion of conscience as a design primitive for social robots. This issue is ultimately of ontological fidelity. I contend that social robots should be modeled after social (human) *beings*, not the poor substitute – *interacting automatons* – of which human beings unfortunately tend to be like.