ON WHAT GROUNDS MIGHT WE HAVE MORAL OBLIGATIONS TO ROBOTS?

(Note: the following thoughts represent only a preliminary sketch on a topic I hope, at some point, to examine more systematically. My primary motive for these thoughts centers around section 3 on Hans Jonas, with whom I had the good fortune to take several courses while I was in the PhD Program at the New School in the late 1960s and early 1970s. While Jonas never offered a course on moral obligations we might have toward smart machines, it seems to me that his treatment of technology, and an ethics for our technological age, offers insights that could be developed and applied to this topic. My remarks in section 3 come primarily from Jonas' course on early modern philosophy (in which Bacon's ideas were extensively covered) and his famous seminar on Heidegger's "Being and Time" (in which being-in-the-world was a central focus). I want to thank my colleague (and good friend!) Herman Tavani who, in several conversations on this topic, urged me to write my thoughts up in a form that allowed at least limited public access.)

There are several ways we might account for any moral obligation humans owe to robots. They all, in one way or another, depend upon conceiving robots such that they have – directly or indirectly – moral status. By "robots" I wish to exclude non-physical software bots, dedicated mechanical devices such as those found on automotive assembly lines, and single-service devices such as automated vacuum cleaners; I wish to include physical devices designed to mimic human behavior, respond to and assist human activity, and that qualify as "smart machines," i.e. they operate by AI programs that run algorithms CS experts typically refer to as "deep learning." In the literature a sub-set of such smart robots are now commonly referred to as "social robots" (SRs). Paraphrasing Kate Darling's (MIT's Media Lab) characterization: a SR is a robot that uses its linguistic and behavioral ability in social ways to interact with humans and elicit human responses. Three fictional examples of a SR would be "David" (the robot boy) in the 2001 Spielberg film *A.I. Artificial Intelligence*, "Ava" (the robot woman) in the 2015 Alex Garland film *Ex Machina,* and the "Terminator" in the classic *Terminator* series of science fiction films. *A* non-fictional example would be the MIT Robotics Lab's robot *Cog*. Let's now assume that SRs *do* have moral status (minimally as moral patients, if not as full moral agents) and that, consequently, humans have moral obligations toward SRs that constrain how humans permit themselves to interact with them. (An example moral obligation might be: humans should not intentionally damage a SR without sufficient reason to do so. I take this to be a relatively uncontroversial robot version of our moral norm: it is wrong for a person intentionally to hurt another human being unnecessarily.) We ask: on what conceptual understanding might such moral status and resulting obligations be based? Here are four possibilities.

1. By analogy with humans. With respect to several central properties (for example, rationality, decision-making, behavioral autonomy, physiognomy) that are widely taken to be central to and even definitive of our humanity, SRs can be considered our lesser analogue. That is: SRs can be taken to have (or to simulate) "lesser," perhaps less general, versions of such human capacities and traits; SRs have (or will have) their own versions of rationality, decision-making, behavioral autonomy, and human-like physiognomy. There is a major school of thought in moral philosophy that argues, in the case of humans, possessing properties that form our humanity provides the basis for humans having moral status and for the moral obligations humans owe to each other. By analogy, then, the presence of lesser (or more dedicated) simulated versions of such properties in SRs should give rise to SRs having (perhaps a lesser form of) moral status, and recognizing their moral status gives rise to moral obligations that humans owe to SRs. This argument is not based on trying to accord robots rights. It is an argument that might be seen as a broadly Kantian approach to the question: moral obligations depend on the moral status of SRs, which in turn depends on their "human-like" capacities of rationality, autonomy, etc.

Problems: like all arguments that use analogy, for two or more things to be analogous they must be similar in enough important respects. The problem is that there can be stronger and weaker similarities; the weaker the similarities, the weaker the argument using analogy. Similarities, in turn, must be evaluated and to accomplish this we must be able to compare the items in question (here: certain properties in humans in comparison with simulated versions of these properties in SRs). We can ask: how similar are human rationality, decision-making, and behavioral autonomy (or whatever essentially human abilities we might choose) to those that are designed to function in SRs? Given that we know much more about the forms these properties have in the case of SRs than we do about the human case, it would seem at this stage that a comparison wouldn't be possible; we just don't know enough about human rationality, decision-making, behavioral autonomy, etc. to allow for a comparison. And without the possibility of comparing these properties in humans and in SRs, estimating the strength of similarity seems not to be possible. Thus, the argument for moral obligations humans owe to SRs that's based on a human-SR analogy must be put "on hold" and cannot be evaluated or accepted until we know more about human rationality, decision-making, behavioral autonomy, etc.

2. By value. SRs clearly possess a variety of values. They not only have economic value, they also embody the creativity, the ingenuity, the knowledge, and the efforts of the people who designed and produced them. As objects, then, some SRs can be appreciated as ingenious creations. SRs also have utility value: they perform a variety of services that benefit people –

improving our well-being and enhancing our lives. Thus, SRs directly contribute, with increasing significance, to human welfare and thus to the human good. Likewise, if SRs have disutility by causing harm to the humans with whom they interact, then they directly contribute to the lessening of the human good. In addition to utility, some SRs might even embody aesthetic value, having been designed with eye toward their pleasing appearance in structure and functioning; their behavior, style, ornamentation, refinement, and proportions are meant to be aesthetically inviting and friendly. Any object that contains such a variety of values should, according to this argument, morally constrain how humans interact with it. People would have a moral obligation, on this basis, not to neglect, abuse, mistreat or damage SRs in ways that would diminish their value. As a great work of art – a painting or a sculpture, for example – (morally) should not have its aesthetic value diminished by intentional damage or neglect (unless there is overriding reason to do so), so too – it might be argued – we have moral obligations not to lessen the values (or increase the disvalues) embodied in SRs unless there is overriding reason to do so. This argument does not attempt to accord SRs rights. It might be seen as a broadly consequentialist (utilitarian) argument: our moral obligations to SRs depend on the ability of SRs to contribute, positively or negatively, to overall human welfare.

Problems: this argument suffers from one of the general problem that consequentialist theories face: namely, it is difficult to form moral norms of obligations when consequences are many and complex. So, for example, if a SR produces benefits for an elderly person by functioning as a caregiver, but in providing this service a human caregiver is harmed by losing employment, what is our moral obligation to this SR? Is it morally right to shut down the SR and give the human caregiver employment, or would this not be our obligation? Even if we could follow out all the proximate and remote consequences for all humans affected by keeping the SR as caregiver (never mind the time and resources this would take), it is doubtful that our moral obligation could be decided. If we now multiply this one case of caregiving by all the chains of consequences of all the SR-human interactions, we see that deciding on the moral obligations humans owe to SRs based on consequentialist considerations seems doomed to chronic incompleteness and under determination.

Another problem with this argument is that it places the formation of moral obligations after the fact, too late as it were. We would like to know our moral obligations with respect to SRs before our interactions with them, so that they might morally guide our interactions with them. However, this argument requires us to interact with SRs first, perhaps in a morally neutral way, and only then – based on weighing the positive and negative outcomes – form our moral obligations.

3. By being-in-the-world. SRs are (or will be) an integral part of the way that human existence takes place. If so, analyzing human existence, especially with respect to technology, should offer insights into a possible basis for moral obligations humans have towards SRs. We find such an analysis in the work of Hans Jonas. Briefly, Jonas' position has two parts: one is epistemic – his analysis of the relation of knowledge with technology, and the other is metaphysical – his analysis of human existence. I will take each in turn.

Jonas sees the Western classical and medieval concepts of knowledge as largely unassociated with technology. During these two great periods, thinkers viewed knowledge as having intrinsic value, to be pursued for its own sake. Contemplating and understanding the universe was valuable on its own or (during the medieval period) perhaps as a way to gain insight into the nature of God. Technology during these periods was seen not so much as a matter of knowledge but more of gaining skill and craft by practice and apprenticeship. Jonas contrasts this view of knowledge with the Modern period, during which a deep redefinition of knowledge took place which he attributes to Francis Bacon. Bacon (according to Jonas) argued that the value of knowledge is not intrinsic, and not pursued for its own sake; rather its value lies in the degree to which knowledge improves the human condition. Knowledge can benefit humanity not as a contemplation of the universe but by "conquering nature"; that is, knowledge of the material world should give us the power to manipulate nature to our advantage, thereby improving human life. In Jonas' view, this redefinition "knowledge is power" unites knowledge and technology; technology is now the power of knowledge, it is knowledge applied to altering nature for human benefit. Since the early modern period, this new notion of knowledge has gained deep cultural roots in the West and has had widespread success in a variety of fields, to the point that today (Jonas believes) humans have amassed sufficient epistemic and technological power to destroy all human life by making the earth uninhabitable (e.g. by thermonuclear war, resource depletion, atmospheric pollution, radical climate change, etc.). Based on his analysis, Jonas calls for a new "ethics of responsibility" as a way morally to counterbalance this vast and potentially dangerous technological power.

With respect to human existence, Jonas was primarily influenced by Heidegger (and to a lesser extent by Whitehead's process metaphysics). In Heidegger's analysis, a person's existence is "being-in-the-world." "Being-in-the-world" does not mean that we have humans (in isolation) and we have the world (in isolation) and then we (conceptually) place humans "in" the world as the space in which human life gets carried out. Rather, being-in-the-world is the very way that human existence happens; it is the fundamental pattern of functioning *on which* the human person is based. Being-in-the-world = human existence, and being-in-the-world is the more primary concept. It is not that a human's life forms the person's world; it is rather that being-in-the-world forms the human that the person eventually becomes. Jonas, a student of Heidegger, took this analysis of human existence as being-in-the-world and developed it in terms of a

"philosophy of the organism." For Jonas, each human is a dynamic unit of activity (i.e. being, existing), a biological organism in a constant process of exchanging material with the natural environment (i.e. the natural world). This (organic) being-in-the-world – as the basic form of human existence – is our reality; in Jonas' view, human nature cannot be correctly conceived without the natural world with which we have constant interaction by way of exchange.

Jonas now combines these two themes; he applied his Heidegger-inspired notion of being-in-the-world to our contemporary technological world. Human being-in-the-world means today human *being-in-the-technological-world*. Technology has not only permeated every last corner of human existence; technology can now be seen as forming humanity just as much as humanity forms technology. This is to conceive of each person's existence as a sort of technologically enhanced network of abilities by which the person interacts with and exchanges material with the natural environment; this "being-in-the-technological-world" conditions, sustains and defines the human person that results.

If we return to our topic of the basis of our moral obligations to SRs within this Jonas framework, it follows that the moral obligations humans have to each other are moral obligations to being-in-the-technological-world. Given that we have moral obligations to being-in-the-t-world, we now ask: What place do SRs have in our technological world? We see that SRs play an increasingly significant part. SRs are perhaps the most significant area in the technological world where digital (i.e. deep learning AI algorithms) and mechanical technologies come together. If this is correct, it follows (within this Jonas framework) that SRs have moral status; that is, they have moral status as technological enhancements of human power. It further follows that we have moral obligations towards SRs. The argument might best be summarized:

1. Humans have moral obligations to each other (premise: whatever the origins of such obligations).

2. Human existence is being-in-the-world (premise: from Heidegger's analysis).

3. Thus, humans have moral obligations to human being-in-the-world (conclusion: from 1, 2).

4. Being-in-the-world is being-in-the-technological-world (premise: from Jonas' analysis).

5. Thus, humans have moral obligations to being-in-the-technological-world (conclusion: from 3, 4).

6. SRs represent an increasingly significant and integral part of the technological world (premise: from empirical observation and prediction).

7. Thus, humans have moral obligations to SRs (conclusion: from 5, 6).

Problems: assuming that the above reconstruction of Jonas' position with respect to SRs is on the right track, there appears to be a tension in his position between, on the one hand, the power over nature humans have through technology (and by implication through the assistance of SRs) and, on the other hand, the relative autonomous parts of technology (especially SRs). This tension is not the same as the problem that Jonas focused on and worried about; namely, that technology has progressed to the point that its activities have outrun the ability of technocrats to anticipate and control them. In both the extent and in the speed with which technological effects happen, technology has achieved a sort of "independence" from human control. So, to take a familiar environmental example, once greenhouse gases reach a certain level of atmospheric pollution, humans will no longer have the means to slow down, stop, or reverse global warming. In this sense, greenhouse pollution (above a certain level) can be said no longer to be within our control, and thus to have gained the status of being an "independent process." And yet, as Jonas notes, technology is collectively a human creation, completely dependent on human existence; if there had been no humans, there would have been no technology (barring the evolution of non-human intelligent life on earth equal to human creativity). Thus, there is a sense in which, within Jonas' analysis, technology is both independent of human control and yet completely dependent on human activity. This tension, however, is not the one I wish to point out.

The problem with which I'm concerned was, to my knowledge, not addressed by Jonas; it's the tension between SRs being autonomous vs. SRs assisting humans in the control of nature. Within the Jonas framework, technology (including SRs) enhances human power to manipulate the world toward the ultimate goal of bettering human life. This view envisions technology as akin to a vast system of "tools" which is at our disposal (even though parts of this set of "tools" might get out of control, as explained above). In this view, SRs, as part of technology, exist to "do our bidding," as it were. And yet, SRs are being designed to be autonomous; that is, their AI programs (their "deep learning algorithms") result, or will soon result, in decisions and behaviors that are completely independent of anyone's interference. One can easily imagine future SRs that are as autonomous as adult humans, even if only in a specific area of activity such as elderly caregiving/companionship, vehicle use, or military drone operation. Being "cognitively" and behaviorally autonomous is a central requirement for SRs having moral status and for humans believing that we are morally constrained in our interaction with them.

It would appear, then, that the Jonas framework for accounting for a SR's moral status, and for our moral obligations with respect to SRs, contains a problem. On the one hand, SRs are part of technology and thus are "on our side," operating along *with* humans in the struggle for our

welfare. Yet, on the other hand, SRs are by design operating independently (i.e. autonomous) of human control.

It is doubtful that the solution to this problem lies in examining what robot "autonomy" amounts to. As noted, it is a restricted or constrained autonomy, not the radical autonomy we typically associate with humans. If, for example, a military drone is designed to be autonomous, its internal decision-making algorithms and consequent behavior would range over options strictly within the field of operations. It couldn't, say, decide to behave in ways that had nothing to do with the military operation at hand; there would be no such options for it to calculate. A human military operator, however, presumably has unrestricted autonomy and could decide to behave in ways that had nothing to do with a particular mission. Would this "restricted autonomy" ease the tension in Jonas' framework? Not really: it would perhaps shift the control/autonomy tension to a more narrow conceptual space, but it wouldn't resolve this tension.

4. There is another approach to the question of our moral obligation to SRs. Kate Darling, a "robot ethicist" working in MIT's Media Lab, has proposed (see her interview on NPR's program Hidden Brain, 1/5/18, 2nd half of program) that moral obligations initiate with *perception*. Darling argues that even if it is a version of anthropomorphizing, as long as people *perceive* SRs as sentient, mindful, intentional, etc. a level of morality emerges such that the people so perceiving feel morally constrained in how they interact with them. This notion of "perception" seems to be a kind of "projection," or a "stance" in the sense of Dennett's intentional stance. According to Darling, such a perception is sufficient to engage/activate moral judgments and beliefs about how the perceivers should and should not relate to SRs. In her experiment, she presented subjects with SRs that had been given names and told the subjects their names. The SRs performed human-like behaviors that the subjects witnessed. The subjects were then given hammers and asked to hit the SRs sufficiently hard to damage them. They refused, claiming it would be *wrong*. It was explained to the subjects that the SRs weren't alive and were, in fact, only smart machines. Still, the subjects refused to damage the robots. In Darling's experiment, the subjects were pressured; the subjects were told that if they didn't damage at least one SR, all would be destroyed. Only one subject agreed to hit a SR with his hammer, all the other subjects continued to refuse (check this fact: was it all the others or a majority of others?). Darling concludes that the relatively brief perception the subjects had of the SRs was sufficient for them to interpret their interaction with the SRs in moral terms; the subjects felt a moral obligation not to damage them even though they had permission to do so, it would be legal to do so, and they would be rewarded as subjects in a university experiment for doing so. (Also,

NPR's On Point of 1/10/18 was partly on robot ethics. Virtual assistants, e.g. Siri, Alexa, were perceived as "feminine" in "character" or as outright "women," while virtual geniuses, e.g. Watson, Deep Blue, were perceived as "masculine" in "character" or as outright "men.")

Posted: 2/2018