Social robots as mirrors of (failed) communion[1]

Niklas TOIVAKAINEN
University of Helsinki

Abstract. The initial point of this paper is that *when* we are engaged with the world, with human beings and morality in a technological or techno-scientific framework, we are projecting an interest of power — our interest to know how things "work" — into our conceptual scheme. In contrast to a technological understanding, I suggest that human beings are characterised by a non-technological moral necessity. I characterise this moral necessity as an inherent responsiveness, an *I-you* relationship, whereas my claim is that a technological conception of morality takes as its ethical basis collective social identities.

Keywords. Technological understanding, moral necessity, social identities, I-you relationship

Prelude

In order to frame this paper I want to begin by taking note of Sherry Turkle's anthropological studies of the perceptions and attitudes towards Artificial Intelligence systems, robotics and digital technologies. In her recent book from 2011 [1], we find convincing signs that people's attitudes towards AI technologies – and especially social robots (and multi-media technologies) – are characterised by

- Narcissistic tendencies to utilise robotics and other technologies to fulfil one's own needs, which in turn is a sign of failed communion with real human beings.
- Robots and AI technologies are ways of distancing oneself from life's sorrows and tragedies, e.g. human mortality.
- Sociable robots are felt to be away of reconnecting to the real world in a world of hyper-connectivity where everyone relates to each other primarily through cell-phones and social medias.
- Sociable robots and technologies more generally are means of "solving" problems stemming from social structures without really addressing the problems: e.g. the idea of introducing robots to elderly care only solves the symptom — that elderly people are increasingly becoming isolated and lonely

[1] This paper was published in *Sociable Robots and the Future of Social Relations*, Seibt. J. *et al.,* IOS Press, Amsterdam.

> — without addressing the problem and the moral difficulty of why elderly people are becoming increasingly isolated and lonely.

These are only some of the examples we find in Turkle’s book. Nevertheless, what it strongly seems to suggest, on an empirical level, is that sociable robots are deeply connected with our desire to avoid and hide from real moral relationships and to devise technologies that effectively help us do this. Obviously, the troubling characteristics that Turkle depicts do not exhaust these technologies and their effects on society. Nevertheless, as it seems to me, they characterise a kind of dark shadow, a suppressed and even repressed dimension of the prevailing technological movement. This paper will interest itself with trying to understand in what sense such anthropological observations re-emerge in conceptual and moral reflections.

Introduction

The notion of sociable robots has increasingly led to what I will call an “ethical turn” in the debate on the nature of AI systems. By this I mean that the question of the “realness” or “genuiness” of AI systems has integrated the question as to what extent AI systems can become genuinely “conscious” with the question as to what extent AI systems can come to display moral agency[2], [3]. In other words, as I would like to see it, the “ethical turn” means that the question of the nature of “the mind”, “consciousness” or “intelligence” has increasingly become to be perceived as interdependent with the question of ethics[4], [5].

Nevertheless, even though I in a certain sense welcome such a turn in the philosophical landscape, I am not all that optimistic about the prospects for new insights arising out of it. My main concern could be characterised the following way: our continuous and blind pursuit for technological (and amongst other things economic) “advancement”, “progress”, or what have you, rests on our willingness to transform and adapt our social reality to accommodate these new technologies and the new social relations, institutions and infrastructures that they give rise to and require. In other words, we appropriate ourselves to meet the requirements of technologies. This applies similarly to ethics. In our pursuit to devise and construct “ethical” technologies (e.g. ethical social robots) we appropriate the language of morality to fit the language of technology, or to be more precise, the language of techno-science. One must obviously keep in mind that this appropriation did not start 10 years ago, or 50 or 100 years ago. Rather, as I would put it, this appropriation has been going on for as long as and proportionally to the extent to which our civilisation has adopted itself to technology (where we draw the line here is not essential, but rather reflects what aspects or features we want to highlight). In saying this, as will become evident throughout the paper, I am not indicating that moral language or morality as such is always and simply a matter of social construction or anything of the sort. As I will try to argue, the appropriation of moral language to meet the requirements of for instance technological or other power interests is always characterised by a dynamic relation to the language of love, or what I will call the moral necessity of human life.

1. Technological understanding of life vs. the moral necessity of human life

In what sense and to what extent can AI systems or robots become moral agents (or alternatively, to what extent and in what sense can robots become “conscious”, “intelligent”, “autonomous” etc.)? For the most part — exceptions included of course — people tend to agree that current technologies, i.e. robots or AI systems, cannot in any serious sense be said to fulfil the criterion of moral agency. But will they perhaps someday reach the threshold of genuine moral agency? On this point opinions seem to become more divided. Some tend to think that already existing technologies clearly indicates that we are getting closer to or are already very close to this threshold, while others take a kind of agnostic position or then bluntly just reject any such possibility. Without at this point taking any stance on the question, I want to explore what is conceptually presupposed *if* one is prone to even entertain the possibility of so called strong (moral) AI.

To begin with, I would claim that the idea of strong AI (even if simply entertained as an agnostic possibility) entails the idea that there is no essential or critical difference between robots and real persons, that is to say, that there is nothing in principle that hinders AI systems from becoming genuinely conscious moral agents. I will characterise such a standpoint as a technological or techno-scientific understanding of phenomena and life. Here is a characteristic account of such a stance: humans and life in general are “*made of* mindless robots [cells] and nothing else, no non-physical, nonrobotic ingredients at all” [6]. Now I want to point out that claiming such an essential unity between nature and artefact obviously goes, so to speak, both ways: machines and artefacts are essentially no different than nature or life, but the main argument and emphasis, I would claim, is really that nature and life are essentially no different from artefacts. The reason for claiming that the latter formulation is the dominant one in a technological understanding of life, is my claim that such an understanding (secretly) perceives the artificial to be that which is *essential.* This understanding was the cornerstone of Francis Bacon’s scientific revolution — which became the cornerstone of modern science as such. In conceptualising knowledge as that which gives man power over phenomena [2][7], Bacon radically challenged the Aristotelian and scholastic understanding of nature as that which creates itself or that which is essential, to an understanding of nature and life as essentially the raw material for mans industriousness. The ontological implication of this was that nature or the natural/genuine is *nothing apart from how man knows it or will someday be able to know it* — and here “knowledge” is conceptualised as that which gives *power* over phenomena[8]. And so Bacon voiced the same conception as Dennett above, namely that the *artificial does not differ from the natural either in form or in essence, but only in the efficient*[3][8].

Much of importance would have to be said about what I have here characterised as a technological understanding of phenomena, nature and being and its socio-economic as well as political and ideological framework. Obviously I cannot go into it here. Nevertheless, there is a specific issue with such a technological understanding that I want to challenge, an issue which has direct bearing on our understanding of AI and sociable robots and will lead us further in our investigation.

[2] c.f. Bacon’s famous formulation “*ipsa scientia potestas est*” or “knowledge itself is power”

[3] Let me also add that by this I am by no means trying to defend or advocate Aristotelianism or scholasticism.

Just now I made the claim that what I have characterised as a technological understanding of nature and being implies or presupposes the idea that nature or phenomena is nothing apart from how man knows or can come to know it. And I added, and will re-emphasise it, that “knowledge” is here specifically conceptualised as an ability to gain manipulative power over phenomena. Without a doubt, sociable robots — as any other AI system or technological devise — is a result of such “knowledge” and its nature is, in a sense, revealed through this “knowledge”. The very straightforward implication of this is of course that these robots are in an important sense derived from *our power* to devise them. That is to say, whether or not robots will become sociable or “ethical” (or whether or not any robots are engineered and built at all!) is in an important sense up to us. This is obviously a very simple, yet essential point.

So one might say that technological devises, for instance robots, do not have any essence independent from how we devise them. The reverse is also true, namely that technological devises (what they are) are not independent of the motivations and aspirations that underlie our will to devise. Rather, as one might put it, they reflect the morally charged dynamics underlying their very creation.

One might surely agree with the notion that an individual human being and perhaps even life as such is not in any strict sense determined by an essence either. Nevertheless, there is something that we cannot avoid, something we cannot escape, so to speak: humans cannot (absolutely/categorically) be unmoved by other human beings, perhaps even other forms of life. In other words, our lives is characterised by our inherent and, one might say, spontaneous or direct responsiveness to others. Surely we can choose to avoid or suppress responding to or acknowledging others, but this will always demand some effort from our side. We might also become so cold hearted that it seems as if we were completely indifferent to others, but this, arguably, is a form of repressing the responsiveness inherent in our relationship to others. Just to take a crude example: If I see a child crying and nobody is there to care for it, I cannot but respond to this cry. Instead of caring for the child, which would be, as one might put it, the direct response, I might turn away from it and argue for myself that someone else is bound to come sooner or later, that it is really the parent’s responsibility to care for it etc. I might also be affected by some terrible trauma involving abandonment leading to a response where I completely blind myself from the existence of the child, or I might have developed some pathological perversion that makes me enjoy watching children suffer etc. The point nevertheless is that whatever happens, the child will move us to respond to it affectively in one way or another. And let us add, when turning away or in failing to respond with care and love, we *must* give reasons for our cruelty[4], whereas when we respond lovingly, with compassion and care (as it were, without any sense of bad conscience) we are not in need of any *reasons* for our actions, for our love and care. — Surely, I might describe the situation and the suffering I saw and say “I cared for the child because it was suffering”, but these are not reasons for *why* the other is significant, why s/he addresses my love and care in the first place.

[4] A paradigmatic example of this would be the slave-owner who reasons to him/herself and others that the slave, although human like, is really sub-human and for example unable to feel pain and sorrow as proper humans do.

Now it is obviously true that we can come to devise a robot that displays *some* characteristic behaviour of compassion and responsiveness. But here there is no necessity to be found in the same sense as with human beings: *we might just as well not devise such robots*. One question which addresses us here is: could life's evolution have *chosen not to make us responsive to others?* My direct answer or claim is: *we humans* design and build robots with specific aspirations, motivations and for specific purposes. Life or life's evolution does not design with any end-specific purposes or motivations. Rather it, so to speak, follows or develops along life's own (specific) nature (be it a "natural" or "super-natural" nature). And, as it seems to be, our responsiveness to each other (perhaps even any living being's inherent responsiveness to other living beings) is an inherent part of our own being — a *moral necessity* — whereas any technological devise does not have any unavoidable nature to it, but is fully dependent on humans to exercise their power. It is in this sense that life is not (at least not essentially) technological and cannot as such be captured by technological or techno-scientific knowledge, despite what Bacon and Dennett claim.

What I have been trying to say is that (i) *when* we are engaged with the world, with human beings and with morality in a technological or techno-scientific framework, we are projecting an interest of power — our interest to know how things "work" — into our conceptual scheme. Or rather, this interest of power creates its own conceptual scheme and appropriates/transforms/modifies for instance the language of morals in order for it to fit our techno-scientific aspirations. The critical point here is: in a technological framework morals (the moral necessity of human life) cannot speak on its own terms[5]. (ii) The other crucial point I have been trying to make is that due to how technology stands in relation to us, the source of meaning in any AI system would eventually come from *us* — although the chain of reasons/meaning might be long and complex. That is to say, whatever "morality" we think we can *devise* will not bear with itself the same inherent moral necessity as that which constitutes our very being, i.e. the being of the very beings that do the devising. The source of morality lies in the very fact that others are always already morally significant to us and this is not something we can choose to have or not to have. Nor is it something "evolution" could have "chosen" to leave out. We, on the other hand, are not predetermined to make robots sociable or ethical, or to devise any robots at all for that matter. To think that an "ethical robot" is unavoidable is to misunderstand what our freedom consists in.

I am of course aware that what I have just now said is a very strong claim and one which hasn't been argued for sufficiently. And this is how it must stand in this paper; as a provocative and challenging claim.

2. Morality is to see beyond social identities

Now *if* my claim that (human) life is not essentially a technological phenomena is true — and that a technological understanding of phenomena suppresses or represses this truth — then the question arises: what is it that makes sociable robots a reality (that

[5] And this concerns not only a technological framework, but any framework or interest that does not let, as it were, the language of love speak for itself.

they are something we can devise) and why are people more and more prone to ascribe emotions to these robots?

In her anthropological studies of people's perceptions and attitudes towards Artificial Intelligence systems, robotics and digital technologies, Sherry Turkle has noted that since the 1970's people's attitudes have moved from what she characterises as "philosophical" or ontological to "pragmatic". That is to say, whereas especially children used to, during the 70's and 80's, pose questions as to whether computers and robots *could be alive*, today they are more prone to think of the "aliveness" of robots in terms of "alive enough for..."[6][1]. Turkle's observation reveals, what to me seems to be a central feature of our understanding of AI systems and especially sociable robots. Namely, as Turkle shows in her studies, the gradual change of attitudes is not simply a question of improved technologies, but much more so a question of the social realities and their effects on interpersonal relationships that are both effected by and in turn effect in their own right the development of robotics and other technologies. That is to say, we should not only look at the technological and theoretical developments, but rather pay more attention to the social reality that is developing and embracing these technologies. This important point was already captured in something Lewis Mumford wrote:

> Descartes, in analyzing the physiology of the human body, remarks that its functioning apart from the guidance of the will does not "appear at all strange to those who are acquainted with the variety of movements performed by the different automata, or moving machines fabricated by human industry…Such persons will look upon this body as a *machine* made by the hand of God". But the opposite process was also true: the *mechanization of human habits prepared the way for mechanical imitations*.[9]

Mumford wrote this in the 1930's in the context of tracing the development of modern western technology and civilisation. Yet the same logic that could be observed as a driving force for both the ideological as well as the practical introduction of mechanisation of human labour and replacement of it by machines in the 17th century is also driving forward the robotic movement. Here is another observation by Turkle:

> Their [the research community and industry behind development within robotics] position (the performance of care is care enough) is made easier by making certain jobs robot ready. If human nursing care is regimented, scripted into machinelike performances, it is easier to accept a robot nurse. If the elderly are tended by underpaid workers who seem to do their jobs by rote, it is not difficult to warm to the idea of a robot orderly. (Similarly, if children are minded at day-care facilities that seem like little more than safe warehouses, the idea of a robot babysitter becomes less troubling). [1]

What both Mumford's and Turkle's observations seem to suggest is that the idea of replacing humans with machines and robots runs parallel with an increasing mechanisation, regimentation and formalisation of social and interpersonal relationships. In fact, what at least Mumford quite explicitly tries to show is that western technology is itself deeply rooted in a desire for personal and societal regimentation and control (coupled together with the appropriate cultural, economic

[6] This phenomenon seems to speak in favour of the predictions of those who hold that the philosophical question of the authenticity of Artificial Intelligence (strong AI) will in the end not be *solved*, but rather *dissolved* through a change of social reality: AI systems become so intimately integrated with human social relations that people stop asking about the authenticity of Artificial systems.

and cosmological context): a will to make life a system that follows exact, predictable and controllable patterns[7] [9].

There is of course something — formative — about our “ordinary” or “everyday” social relationships that is in a broad sense inherently, as one might put it, technological. By this I simply mean to say that there is something that characterises our social reality which is built on a demand for regimentation, control, formality and domination. I am of course here thinking of our social/collective normative identities that serve the function of regulating interpersonal relationships. And, as I will now shortly try to argue, it is exactly because of this trait that the idea of a formal system (a robot) potentially being a candidate of moral agency is attractive. For as I would suggest, the *rules* and *principles* guiding our collective normative identities are something formal systems can, in principle, come to simulate. Or to be more precise, it is because we are tempted to think of morality as a system of norms and principles that it seems to be an open possibility to devise a formal system that would be able to follow/understand and internalise them.

But, one may ask, aren’t all our relationships deep down formal? In other words, is there something about our relationship to others that is a-formal, something that is not determined by our social identities — by our social norms and conventions — something that is not shaped by principles, rules or mechanisms? My suggestion here is that this is first and foremost a *moral* question, a question that needs to be answered in moral terms; it is a question that involves a constant moral challenge and self-investigation, a clarification of how much of our relationship to others is determined or formed by the formal categories of e.g. “parent”, “child”, “man”, “woman”, European”, “Indian”, “worker”, “politician” etc. as they are understood in terms of collective normativity, and to what extent one is open to the other as an *I* to a *you?*

Now the important thing I want to stress here is that the formal or normative aspects of human social and interpersonal relationships cannot really be understood as exclusive or fundamental; as the source of our morality. As I have been stressing earlier, one of the essential things that makes us human is that we cannot but respond to or be moved by others: the other’s presence will always address and claim, as it were, a response from us, a response which we might try to avoid, control or repress but never annihilate. This is the source of our moral being and the source of any ethical principle or theory [10], [11]. That is to say, it is only because of this moral necessity that ethical theories have any sense. For, as I would claim, normative ethics is not what morals is about, but rather makes itself known only insofar as it is a response to the difficulty and failure of embracing or opening oneself to the possibility of responding to others with or in love: ethical principles, as important and necessary as they might be, are always signs of a failure of communion and love, since their primary function is to manage the tension that stems from peoples failures to live in openness and love. One might also say that ethics is an endeavour requiring rational or intellectual capacity by which one mediates and manages moral difficulties, while the unavoidable affective responsiveness to others is neither intellectual nor emotional, but the very thing that

[7] One of Mumford’s leading observations is that western technology and science builds essentially on the reinvention and perfection of the mechanical clock (time keeping).

gives morals its sense. Ethics (as a normative system) is, as the Danish philosopher K.E. Løgstrup has put it, always a second best [12].

Seen from another angle, the issue might be characterised as follows: As we are all quite aware, what constitutes any given culture's social norms and values can be, and usually are, loaded with morally corrupt or suspect conceptions. Historically one of the main traits of such morally questionable conceptions present in most cultures at some point of history has been the idea that being a moral patient applies only to a specific and exclusive group of people. So for instance, as it was equally perceived in collective social normative terms by both the ancient Greeks as well as by slave-owners in the southern states, slaves were perceived to be sub-human and thus were not subject to the same moral dignity as the prevailing class. The question to be asked is of course from what source do people draw moral insight as to what is problematic with the social norms? Obviously the insight cannot come from the normative structure itself. One might of course say that there is some intellectual insight into some universal ethical principle, and in a sense this might be so. Nevertheless, the principle must in some way both relate and be true to what I have called the moral necessity of human life. In other words, moral insight is always a question of *realisation*. Take the case of the slave and the slave-owner. The insight, even though it might be conceptualised (by the slave-owner) as an insight into some supposed universal ethical principle (e.g. the equality of every human being), does not constitute or found the moral relationship between the salve-owner and the slave. Rather, when it becomes apparent to the slave owner that her/his relationship to the slave is morally corrupt, the slave owner *realises* that s/he has always been morally related to the slave, albeit in a repressed manner. The universal principle might surely be said to do the job of regimenting and controlling the (by now hopefully former) slave-owner's moral responses to those who before were perceived as sub-human, of managing her/his moral corruption. But intellectual (or emotional) insight cannot found or ground the *sense* of morality. That we are able to see beyond our collective social norms depends on our moral energy to open up ourselves (more) truthfully and in a sense courageously to the possibility of responding to the other lovingly or *in* love. Or, one might say, a moral insight is to understand (more) openly and truthfully how we really experience others, what significance others have and how inescapable to us they are [13]. So as the source of our morality lies in our inherent responsiveness to each other it is something that "cannot be explained or inferred from anything else" [13], thus something a-formal and something non-devisable, i.e. not a result of our power over phenomena.

To summon up, the idea of strong AI is, as one might put it, the flip side of the idea that one's moral relationship to others is something mediated exclusively *via* collective social (formal) identities[8].

3. Conclusion

[8] I would like to note here that when I suggest that there is something like an *I* meeting a *you*, as opposed to individuals meeting each other exclusively via their social identities, I am not saying that in such genuine encounters social identities would be, as it were, non-existent. The point is that in a genuine encounter I do not let the social identities determine the significance or meaning that the other person has for me. Rather, as it seems to me, it is the other way around: it is the encounter between persons that give social identities significance.

At some point I wrote that: “One might say that technological devises, for instance robots, do not have any essence independent from how we devise them. The reverse is also true, namely that technological devises (what they are) are not independent of the motivations and aspirations that underlie the will to devise. Rather, as one might put it, they reflect the morally charged dynamics underlying their very creation”.

Going back to Turkle’s observations I alluded to in the *Prelude*, I would suggest that it is this straightforward conceptual point that is reflected in the distressing reality she describes. So what are our motivations for creating sociable robots? On the empirical level Turkle’s studies indicate that sociable robots are deeply connected with our desire to avoid and hide from real moral relationships and to devise technologies that effectively help us do this — in other words, to uphold a culture that is more and more making human relations controllable and replaceable. As Turkle writes: “The idea of an attentive machine provides the *fantasy* that we may escape from each other” [1]. What I have been trying to say, on a conceptual level, is that this desire to control and escape each other — to control and to escape the inherent responsiveness we have toward each other — is not uniquely a characteristic of AI technologies. Rather, one could say that AI technologies is both related to and in a certain sense even an extension of the age old human/animal desire to control and regiment interpersonal relationships through techniques such as social norms and values. Understanding one’s relationship to others exclusively *via* social norms and identities reduces or rather diminishes and represses oneself and the other to a purely social mechanism.

References

[1] Turkle, Sherry, *Alone Together*. Basic Books, New York, 2011.

[2] Sparrow, Rob, *Can Machines be People? Reflections on the Turing Triage Test*. In Lin, Abney & Bekey (eds.) *Robot Ethics*, MIT Press, Cambridge, Massachusetts, 2012.

[3] Wallach, Wendell & Allen, Colin, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, New York, 2009.

[4] Anderson, David, Leech, *Machine Intentionality, the Moral Status of Machines, and the Composition Problem*. In Müller, V.C. (Ed.) *Philosophy and Theory of Artificial Intelligence*, Springer-Verlag, Berlin, Heidelberg, 2013.

[5] Gerdes, Anne, *Ethical Issues in Human Robot Interaction*. In Nykänen, Riis & Zeller (eds.) *Theoretical and Applied Ethics*, Aalborg University Press, Aalborg, 2013.

[6] Dennett, Daniel, *Sweet Dreams*, MIT Press, Cambridge, Massachusetts, 2006.

[7] Bacon, Francis, *Novum Organum*, Bottom of the Hill Publishing, Memphis, 2012.

[8] Proctor, Robert, *Value Free Science?*, Harvard University Press, Cambridge, Massachusetts, 1991.

[9] Mumford, Lewis *Technics & Civilization*, fourth edition with a new foreword by Langdon Winner, University of Chicago Press, Chicago, 2010.

[10] Backström, Joel, *The fear of openness*. Åbo University Press, Åbo, 2007.

[11] Nykänen, Hannes, *Morals*. In Nykänen, Riis & Zeller (eds.) *Theoretical and Applied Ethics*, Aalborg University Press, Aalborg, 2013.

[12] Løgstrup, K. E., *The Ethical Demand*, University of Notre Dame Press, Notre Dame, 1997.

[13] Westerlund, Fredrik, *Heidegger and the problem of Phenomenality*, Philosophical Studies from the University of Helsinki, Unigrafia, Helsinki, 2014.