CSS
ETH Zurich

4 April 2018

Artificial Intelligent Agents: Prerequisites for Rights and Dignity

By Nayef Al-Rodhan

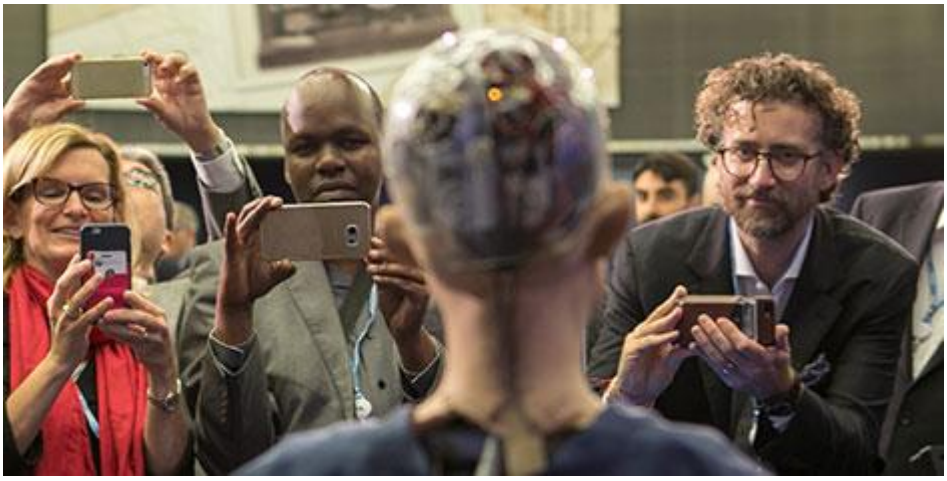

Image courtesy of ITU Pictures/Flickr. (CC BY 2.0)

This article was originally published in Volume 2, Issue 2 of Age of Robots magazine on 6 March 2018.

IBM's Deep Blue was a chess-playing computer that achieved remarkable success in 1997 when it defeated the world champion Gary Kasparov in 19 moves. Kasparov had never lost a match to a human in under 20 moves. He managed to beat Deep Blue in the next games but was again defeated the following year after Deep Blue received an upgrade—and the unofficial nickname "Deeper Blue". This was a landmark moment in artificial intelligence, but at no point was the genius chess machine deemed worthy of "rights". Although theoretically able to visualize 200 million chess positions per second, Deep Blue had limited general abilities and could not work on other tasks beyond what it was programmed to do—such as playing chess, in this case.

However, when robots started speaking and interacting with humans—even creating art and more sophisticated forms of expression, including music—it became clear that their functions were not only multiplying but also starting to rival human capacities.

In 2015, Google's AlphaGo beat a top human player at a game of Go, an ancient Eastern strategy game that involves competencies deemed uniquely human (e.g., intuition). More recently, in a paper in *Nature*, researchers at DeepMind explained how AlphaGo received an upgrade with an algorithm based on reinforcement learning, which allows the computer to learn by itself, without human input, and effectively become its own teacher.

Increasingly, the question is no longer cast in terms of robots' instrumentality but of robots as peers to humans, eventually deserving rights and dignity. In just a decade, human responses to robots have ranged from curiosity and amusement (seeing robots as smart gadgets), to wariness and alarm, and

more recently, as entities deserving citizenship. Saudi Arabia's granting of citizenship to Sophia, a humanoid robot, may be an honorary and symbolic title, but it nevertheless sparks new questions about human relationships with intelligent robots.

Humans and Robots: From Paternalism to Egalitarianism

The current predicament around granting rights to robots is preceded by a long debate on human–robot interactions that began half a century ago. In the interim, some fascinating theories and anthropological studies have described this ambivalent relationship.

One of the earliest hypotheses was put forward in 1970 by Japanese robotics professor, Masahiro Mori, who proposed the concept of the "Uncanny Valley" to describe the interaction between humans and android, humanlike robots. According to this theory, imperfect-looking humanoid objects, which appear similar and yet different replicas of human appearance, will provoke dislike and a strange feeling of revulsion in human observers. Some recent experiments testing the validity of the Uncanny Valley theory have shown that on a spectrum of robot appearances, as faces started to appear more human than mechanical, they were perceived as more unlikable; however, when robot faces started to appear nearly human, likeability increased sharply, but in a precarious way, as even minor faults would disturb social interaction.

Although refuted by some, the Uncanny Valley theory has encouraged some provocative hypotheses about the way interactions and trust can best be fostered in human–robot relations.

Additional joint research from universities in Japan complements some of the premises of the Uncanny Valley thesis, showing that humans can go as far as to empathize with robots in "pain", almost to the same extent as humans in pain. On further investigation, however, the researchers found that the EEG scans of the study participants showed weaker intensity of brain potentials in their responses to the pain of robots than to that of humans. (It is not clear how that sense of empathy could be expressed for robots that do not have a humanoid nature.)

This research would suggest that the more we are able to anthropomorphize robots, the more we could imagine we might be able to understand their perspectives and thus build trust with them or feel empathy.

It is likely this explanation of top-down empathy is in fact a consequence of a more complex set of factors that shape our relationships with robots. MIT researcher Kate Darling has suggested that social bonds with robots are due largely to three factors: physicality (when robots exist in *our space*, not on a screen), perceived autonomy of movement, and social behavior (they can communicate with humans). In an experiment she conducted with other colleagues, the participants were required to beat a group of small robots to death. The aversion to abusing the robots was clear. As humans, we know rationally that robots do not have intrinsic dignity, but we may feel empathy for them because we see in them some reflection of ourselves and thus some of our fears, such as the fear of feeling pain. In other words, a robot cannot evoke the merely neutral feelings of annoyance one might have when the kitchen toaster breaks down.

The unique characteristics of robots have reinforced the idea that they deserve a different kind of approach. Additionally, as machine learning progresses and with it the intelligence and autonomy of robots around us, the questions surrounding the responsibilities of robots will only become more

complicated. One way to address this issue has been to compare it to animal rights, which philosophers and ethicists have deliberated for centuries.

Should we extend rights to robots modeling a similar logic that is used for animal rights, for example? Some philosophical questions emerge instantly, such as: should humans react as principled and responsible guardians? But this thinking is largely outdated, even in relation to animal rights. It presupposes the idea that humans are superior masters who are qualified and able to make decisions about who lives and who must die.

Nonetheless, the parallel to animal rights still holds some value in that the granting of rights and assigning legal personality comes with a desire to connect other agents to humankind.

Although we know that robots do not feel pain and do not have consciousness, yet—the comparison with animal rights is relevant in that it prepares us for a future where we would be more closely connected to AI agents.

Moral Robots: Rights Must Be Connected to Morality

Robots do not have a biological life, or feelings, or the ability to reproduce. They can teach themselves how to play chess but cannot be self-sustaining outside of the will and effort of human manufacturers and engineers. This in itself is a strong enough reason to justify why currently we still think of robots as mechanical slaves, not as entities in their own right.

This thinking is also reflected in legislation that covers liability in this field. In the EU, legal liability for harm caused by robots falls on the manufacturer and the foreseeable damage derived from any manufacturing defects. Increasingly, however, the EU is starting to recognize that when it comes to robots, a new approach to personhood will be needed in the not-so-distant future. Should robots become complex to the point where they can make moral decisions instantaneously, including life and death decisions, a new notion of "electronic person" might be more appropriate.

These ideas were elaborated in a study for the JURI Committee of the European Parliament entitled "European Civil Law Rules in Robotics", in which it was noted that the increasing presence of autonomous robots will create a split in societal values, one so profound that it cannot be matched against the disruptions caused by the Internet and digital technologies. It will be transformative in an existential sense. The authors of this study advised extreme caution: going an extra step and blurring the line "between the living and the inert" would shatter Europe's humanist foundations. It would be wrong, the study concluded, to assign person status to a robot because this would demote mankind. A robot's only purpose should be to serve humanity.

Nevertheless, a discussion of rights will become urgent if the technology progresses according to Moore's Law. At the moment, we can still choose to debate the rights of robots, but that freedom of deliberation might not be there in a few decades.

It is hard to pinpoint when that moment will come (and many scenarios are based on unrealistic predictions), but it seems safe to suggest that later in the century it will be clear enough where the technology is headed.

The truly existential questions will appear when the technology allows for a transition from a top-down morality (meaning that the programmer inputs moral values in the robot) to a bottom-up morality, whereby robots can learn moral competencies through a socialization process, in their environments, similar to how humans achieve a moral compass. This will also mean they are outside of human tutelage and unpredictably free in their acquisition of moral values. What currently separates us from that moment is the technological ability that would permit such learning by robots.

Recent advances in neuromorphic technology promise to bring humanity ever closer to that prospect, however. Neuromorphic (or "brain-like") chips aim to emulate the neural architecture of the brain and the unique way in which neurons work through intricate connections, and nonlinearly, as opposed to standard 0-1 computing. Robots equipped with such chips could learn and process information in a similar fashion to humans.

The consequences of this technology could be a turning point in the debate on robot morality and rights. Not only will robots learn like humans, but their moral compass will develop in an interactive manner by virtue of their social cooperation with human counterparts. Whose morality will they inherit and whose values will they adhere to? I previously theorized that humans function as a predisposed *tabula rasa*, that is, our values are shaped through upbringing and environment, but we are also hardwired in a fundamental way, and that is our instinct for survival. Robots that learn by doing will be similarly influenced by their environment and might even acquire an instinct for self-preservation, even at the expense of humans. A system of rights and responsibilities could be indispensable then to prevent a state-of-nature situation.

Rights and Dignity

For Thomas Hobbes, rights were crucial to life itself: a life without legal rights was one spent solitarily and in danger. For robots to acquire rights they would need to fear for their lives (i.e., have consciousness and self-awareness) and need protection from other hostile entities, human or AI. That can only happen when the technology advances to the point that robots do not have a human operator behind them. Furthermore, if robots were to acquire both the intelligence and the means to cause harm and destabilize the social and political order, they would effectively be living in a state of nature within the Leviathan—that is, within the systems of sovereignty and laws that undergird modern states but which currently only bind humans. Then we would need to think of robots as legal actors, legally responsible for their actions.

At this stage it is impossible to build a strong case for robot rights. So far, nothing indicates that robots deserve rights and legal recognition as separate entities, but that is only because the technology has not yet developed enough to make them fully autonomous. According to Kant, rights arise from rational will, and robots do not have that rational *will*, even though they can make perfectly logical conclusions and associations.

Quantifying the exact skills that would be needed to give rights to robots is a difficult task, and compiling a list of criteria would also be problematic: humans are not uniformly similar in their abilities and still possess rights. We could perhaps resolve the dispute by thinking about such criteria in more general, neuro-philosophical terms. With insights from neuroscience, I previously proposed a theory of human nature as "emotional amoral egoistic". What gives unique distinction to our species

is: first, our emotionality—we are more emotional than we think we are, and our emotions are involved in our decision-making; second, our amorality—we are neither moral, nor immoral; our moral compass is constructed in the course of existence; and third, our egoism—we are predisposed in one fundamental way, in our search for survival, which is a basic form of egoism. AI agents would need to possess similar prerequisites in order to deserve rights: first, a capacity to feel and display emotions; second, a capacity for moral choices and accountability; and third, a capacity for self-awareness and personal egoism.

If these prerequisites were met, AI agents would have emotional amoral egoistic attributes, which now only humans possess. This would also mean that societies would have to guarantee them the nine dignity needs that I have advocated before, which are critical for inclusive good governance, namely: reason, security, human rights, accountability, transparency, justice, opportunity, innovation, and inclusiveness.

These nine dignity needs are essential in order to limit the tensions between the emotional amoral egoistic attributes of human-like robot nature, and ensure sustainable and stable future societies.

The debate remains contentious, however, because it goes at the heart of what makes humans worthy of rights in the first place. Surely there are other entities that have legal personhood, but are not human beings—corporations in the United States, for example, or some natural entities in India? But granting robots rights would be different because robots (unlike corporations or rivers) have intelligence and social skills, and the implications are far more profound.

What is needed to ensure we do not reach that stage, which would lead to a collapse of our civilization, is to advocate for responsible research and a thorough ethical check on the development of artificial intelligence.

About the Author

Prof. Nayef Al-Rodhan *is an Honorary Fellow at St Antony's College, University of Oxford, and Senior Fellow and Head of the Geopolitics and Global Futures Programme at the Geneva Centre for Security Policy. Author of: The Politics of Emerging Strategic Technologies. Implications for Geopolitics, Human Enhancement and Human Dignity.*