Artificial Intelligence

The Shylock Syndrome

DAVID R. LAWRENCE, CÉSAR PALACIOS-GONZÁLEZ, and JOHN HARRIS

Abstract: It seems natural to think that the same prudential and ethical reasons for mutual respect and tolerance that one has vis-à-vis other human persons would hold toward newly encountered paradigmatic but nonhuman biological persons. One also tends to think that they would have similar reasons for treating we humans as creatures that count morally in our own right. This line of thought transcends biological boundaries—namely, with regard to artificially (super)intelligent persons—but is this a safe assumption? The issue concerns *ultimate moral significance*: the significance possessed by human persons, persons from other planets, and hypothetical nonorganic persons in the form of artificial intelligence (AI). This article investigates why our possible relations to AI persons could be more complicated than they first might appear, given that they might possess a radically different nature to us, to the point that civilized or peaceful coexistence in a determinate geographical space could be impossible to achieve.

Keywords: artificial intelligence; humanity; motivation; survival; moral significance; personhood; nature

> SALERIO Why I am sure if he forfeit, thou wilt not take his flesh,—what's that good for?
>
> SHYLOCK To bait fish withal, if it will feed nothing else, it will feed my revenge. He hath disgraced me, and hindered me half a million; laughed at my losses, mocked at my gains, scorned my nation, thwarted my bargains, cooled my friends, heated mine enemies, and what's his reason? I am a Jew. Hath not a Jew eyes? hath not a Jew hands, organs, dimensions, senses, affections, passions? fed with the same food, hurt with the same weapons, subject to the same diseases, healed by the same means, warmed and cooled by the same winter and summer, as a Christian is?—If you prick us, do we not bleed? if you tickle us, do we not laugh? if you poison us, do we not die? and if you wrong us, shall we not revenge?—If we are like you in the rest, we will resemble you in that. If a Jew wrong a Christian, what is his humility? Revenge. If a Christian wrong a Jew, what should his sufferance be by Christian example? why, revenge. The villany you teach me, I will execute, and it shall go hard but I will better the instruction.
>
> William Shakespeare, *The Merchant of Venice*, Act III, Scene 1[1]

Introduction

When we imagine the human race encountering paradigmatic but nonhuman biological persons—meaning "a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing in different

Cambridge Quarterly of Healthcare Ethics (2016), **25**, 250–261.

doi:10.1017/S0963180115000559

times and places"[2] (e.g., Neanderthals, Spock, or ET)—we (almost) automatically think that the same prudential and ethical reasons for mutual respect and tolerance that we have vis-à-vis other human persons would hold toward them. We also tend to think that they *would have* (and certainly, if we know what's good for us, that they *should have*) similar prudential and ethical reasons for treating us as creatures that count morally in our own right.

In fact, this line of thought transcends biological boundaries in that we also are tempted to assume that we would be morally bound to treat artificially intelligent persons (or artificially superintelligent persons) as we treat human persons, irrespective of whether or not these creatures have been created by humans, were encountered in outer space, or turned up here in their spacecraft. In this respect most people reject, at least in principle, what might be called bioism: the prejudice or bias in favor of biological entities with *X* interests and capacities over those of nonbiological entities with comparable *X* interests and capacities.[3] This means that we hold that the same moral wrong would be committed if someone were to kill an innocent human person or an innocent artificially intelligent person. This issue of course is how we "flesh out" (and we use that term deliberately) what justifies thinking of an innocent artificial intelligence (AI) as also a person. The issue in short concerns what might be termed *ultimate moral significance*—that is, the significance possessed by human persons, persons from other planets, and nonorganic persons in the form of AI if and when they appear.[4]

One of the present authors, John Harris, wrote about AI in 1985 and started to think seriously about how we, humans, and they, AI creatures, might react to one another. At that time Harris explored the possibility of extraterrestrial AIs and suggested that

> the question of whether or not there are people on other planets is a real one. If there are, we need not expect them to be human people (it would be bizarre if they were!), nor need we expect them to look or sound or smell (or anything else) like us. They might not even be organic, but might perhaps reproduce by mechanical construction rather than by genetic reproduction.[5]

He then went on to speculate that if their technology proved to be superior to ours (perhaps the proof of superior technology would be them turning up on, or in near proximity to, the earth rather than us tracking them down in some other galaxy), it would be of paramount importance for us to convince them that we are also persons, if not just like them, at least enough like them to matter—in short, that we are persons with whom they would rather have lunch, than have for lunch.

Now, even when we maintain, in principle, that a symmetrical moral relation should hold (i.e., each party treats the other according to its moral status) between human persons and AI persons, two considerations come to mind that might lead us to think that our relations with them could be way more complicated than we usually make believe they could or should be. The first reason is that the creatures (created or encountered) might possess a radically different nature to us, to the point that civilized or even peaceful coexistence in a determinate geographical space would be impossible to achieve. This might be due, for example, to the impossibility of setting reliable limits to the *aims and purposes* of a human-created AI person. This would apply particularly to an AI capable of

thinking about its aims and purposes and adapting itself in ways not envisaged by its designers and over which they have no effective control, just as is true to an extent of us organic, ape-descended humans concerning one another. The second reason is an epistemological one. We might not realize that we have created (or encountered) an AI person. If this were to happen, we would risk not treating "her" (perhaps AIs, like ships, are conventionally female?) as morality requires that she/it be treated.

Here, we focus on the first reason. We investigate why our possible relations to AI persons could be more complicated than at first might appear due to issues surrounding the AI's nature. Let's start by saying that the control problem (i.e., how to regulate, or effectively influence, other beings in such way that we are not put in harm's way by their actions or inactions) that we would have when dealing with AI persons, or when thinking about creating AI persons, is not the very same problem, or at least not precisely the same problem, we humans have with one another. It is not so, because the answers to the following questions vary depending on whether we are talking about a human person or an AI person: How can we minimize the risk posed by people whose actions or plans threaten other people or the planet? How can we eliminate or mitigate the risk posed deliberately or accidentally by other people through wickedness, negligence, insensitivity, stupidity, or *superintelligence*? How can we stop the proverbial village idiot[6] or the village genius from destroying the global village, or how can we stop the agent who fails in his or her or its duty to act for the best "all things considered" from doing likewise?[7]

One obvious answer when dealing with AI persons would be to try to motivate them just as we try to motivate humans, first by educating them, showing them that some sorts of beings are intrinsically valuable, offering the AI rewards, or threatening them with punishment. The problem with this solution is that it is too parochial and almost certainly doomed to fail.[8] In his book *Superintelligence: Paths, Dangers, Strategies*, Nick Bostrom rightly warns against anthropomorphizing the capabilities or motivations of AIs, or superintelligent AIs. This worry is warranted by the fact that most usually we imagine (and in certain cases believe) that other creatures (e.g., aliens or AIs) possess human minds, and thus that they respond to stimuli as such. It is not that we think that they literally possess human minds (with the peculiarities of our evolutionary history) or human brains (the physical basis for human minds) inside robotic bodies or super computers. What happens is that we assume, perhaps unreflectively, that these creatures are motivated to act by the same types of considerations that motivate action in humans (i.e., that we have overlapping or congruent interests that motivate us) and also that they are demotivated by the same sorts of things that demotivate us.

When we make believe that we encounter aliens and AIs, often we imagine the proverbial wolf in sheep's clothing (i.e., nonhumans passing as humans for their advantage). But we encounter such wolves because we have made an epistemological mistake. Although terrestrial or extraterrestrial biological organisms (if there are any) are likely to share certain motivations, if in fact they arose from similar evolutionary processes, human-designed AIs, by contrast, might not share any of these motivations. This can be the case because intelligence and goals are not linked in a specific and necessary way, much less in a way that allows biological beings like us to survive and thrive. It is from this nonrelation between intelligence and goals that Bostrom proposes the orthogonality thesis (OT): "Intelligence

and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal."[9]

A characteristic of the OT is that it does not require us to say anything about rationality or reason. The OT is specified in terms of intelligence, which Bostrom understands as "something like skill at prediction, planning, and means-ends reasoning in general."[10] For we humans, the existential danger that AIs could most certainly impose derives from the fact that AIs could have goals that are incompatible with our survival but compatible with, and perhaps necessary for, the survival or the achievement of the goals of the AI (we do not address in this article issues that would arise with an AI who is reckless or careless of its own survival or the survival of its kind). What makes this even more worrying is that many of the goals an AI could have may be contradictory to the requirements for humans to survive and thrive, which are in any event highly difficult to meet. Why is this the case? Because of all the possible goals that there could be, only a minimal fraction of them are likely to be congruent with the survival of the earth as we know it, and an even smaller number are compatible with the survival of humankind or even of posthumankind.

Motivation

What might a machine life- (or existence-)form or a silicate being actually require to survive and thrive? What needs might drive an artificial intelligence to act toward self-fulfillment? For our present purpose, we can perhaps discount simple programmed commands, instead focusing on AIs with at least a measure of autonomy in their actions. As discussed elsewhere in this article, we cannot assume that these aims and goals would match our own or even be intelligible to us ape-descended creatures of flesh and blood.

One goal it might be reasonable to assume might be held by an AI would be the continued existence of the being and/or its kind. This, it could be argued, is the purpose of the seven commonly accepted human "life processes,"[11] or indeed that of the more nuanced academically accepted physiological functions of life—namely, homeostasis, cellular organization, metabolism, growth, evolutionary adaptation, stimuli response, and reproduction.[12] In *Homo sapiens* and indeed most complex organisms, a lack of any of these characteristics would prohibit life, either by failing to support the organism or by leaving it completely vulnerable to outside hazard. Even single-celled organisms—prokaryotes, eukaryotes, and archaea—are each subject to the majority of these processes, and borderline cases—"organisms at the edge of life"[13] such as viruses, which do not conform to so many commonly recognized key markers of life[14] that they might be considered as simply being organic chemical structures—are subject to at least one.

This latter, ubiquitous function is, of course, reproduction—be it by sexual reproduction or even self-replication, as in a virus or other cellular structures. Many attempts have been made to define life, or to distil it to its essence, and one exhaustive review and analysis, by Trifonov, concludes simply that "life is self-reproduction with variations."[15] It is a reasonably rational assumption to make that any novel or at least newly discovered form of life would follow this pattern and possess, if nothing else we might understand, an aim or at least a propensity to propagate and thereby, pace Richard Dawkins, serve the interest of its genes or their equivalent.[16]

Empirical work, limited as it may be at this time, appears to bear this out. Hod Lipson and colleagues at Cornell demonstrated the "spontaneous emergence of self-replicating structures" in a simulated group of simple, undirected automata[17] without any selection or extraneous reward for using the trait; the "molecubes" exhibited a distinct tendency toward self-organization and the replication of these structures, with populations among different groups fluctuating in relation to one another. As one populace waxes, another wanes. Although this latter activity cannot properly be called hostile competition, given the lack of an organizing central "mind," it is nonetheless an intriguing microcosm of the very concerns we hope to address in this article and is a point to which we will return shortly.

Lipson's work is interesting in that it implies that some form of Trifonov's determination of life applies to a case in silico. Had Lipson's group sufficient resources to build the requisite (prohibitively large, hence the simulation) number of physical, mechanical molecubes[18] and set them loose, it appears probable that the same behavior would have been observed. Molecubes, physical or not, are comparatively uncomplicated, and the simulation only operates within certain parameters. Although they do spontaneously propagate their numbers, it is difficult to say whether this is in service of some inherent drive—let alone goal—to survive or merely the natural expression of the exercise of the molecubes' limited abilities. It could be argued that this distinction is unimportant—a pathogenic virus does not proliferate with intent but rather carries out the functions and processes it is capable of performing. Either way, the virus acts in its own—unconscious—interests, primitively understood. We could draw the conclusion that Lipson's machines exhibit the fundamentals of life in the same manner as a virus and, by extension, necessarily share their "goal" of maintaining it.

Immortality

Given that we humans share this fundamental goal of survival (though we have the capacity to choose to ignore it in favor of other interests[19]) with "lower" orders of beings,[20] it stands to reason that an artificial super- or human-commensurate intelligence would also be subject to it, with the same caveat. We must be wary when drawing this comparison, though: there is a significant difference, beyond substrate, between man and machine. *Homo sapiens* and almost all other known species are (at present) senescent and fleeting—despite any individual or collective survival goal, we wither and die.[21] To achieve the latter in lieu of the former we reproduce. Our AI (perhaps) compatriot, however, is not necessarily subject to the same weakness. It may be functionally immortal and therefore not subject to the same drive to proliferate as are we—that is, so long as it has not given itself the sensual satisfactions (or their nonorganic equivalent if there is one) of the Greek immortals, to have sex and procreate with humans and with each other. It is able to survive—fulfilling that most basic of aims—indefinitely, without any need for a line of descendants to keep its kind "alive." This is problematic—why might the molecubes self-replicate as they do if it is unnecessary?

Perhaps a sole, individual AI would be content to exist alone, secure in the knowledge that it is surviving (assuming it wishes to do so). This would, however, require that it disregard outside hazards and resource requirements. Once these are taken into account, the AI would be in a much less secure position in regards to achieving its basic goal.[22] Presumably, then, it would act accordingly in pursuit of

that goal. Any form of AI, be it ensconced in an individual, physical shell such as an android or one that exists more ephemerally within a wide digital network, would require energy to continue to "survive" and operate, just as we humans die without appropriate nutrition and sunlight. An AI would require complex component resources—or the means to manufacture these—for repair, much as do our bodies. The laws of physics decree that there is a finite quantity of each resource—however vast—available, and it is here that we meet the rub. These resources must be harvested, and the history of mankind is nothing if not evidence for the destructive competition engendered by groups attempting to harvest even plentiful or renewable resources in their own interests.

Here, we might think back to the interesting behavior observed in Lipson's robots. Their subpopulations wax and wane as they compete for resources (in this case, loose molecubes combine into replicated structures). To maximize chances of survival over group B, group A might proliferate in order to maximize its opportunities. Similar behavior is familiar to us from the animal kingdom, for instance, with small animals reproducing in large quantities to overcome the rate of attrition. It perhaps follows, then, that a finite availability of resources would engender reproductive behavior in an AI. In the long term, its motivation to survive and ours would likely be incompatible.

If the intelligence relationship between an AI and the molecube is similar to that between us and a bacterium, it is fair to say that the AI is likely to have rather more discretion in its actions than the Cornell robots. If it is possessed of a moral faculty, it might judge us worthy of conservation, as we do for other species. It may choose to limit its population at some ideal number. Alternatively, it might simply be rational, at least in the short to midterm, for an AI machine being to develop the goal of ensuring *our* survival. But we perhaps should not count on this.

In the field of space exploration, there has been much thought devoted to solving the problem of how we might send probes across interstellar distances. Many of these are variations on the Von Neumann machine concept,[23] in which a machine gathers raw materials during its journey and gradually constructs the necessary industrial infrastructure to produce a replica of itself, which then travels on to do the same, and so on—thus covering vast areas of space. However, as Freitas calculates in an extremely detailed blueprint for such an enterprise, to create this infrastructure would take a large variety of task-focused robots (for instance, atmospheric miners, excavators, metallurgists, chemists, fabricators, quality assurance, power plants, etc.), and at least 500 years from planetfall.[24] Similarly, if we were to turn our AI loose into the world without access to our existing industrial complex, its generation time would be somewhat uncompetitive. Given access, the AI would, we must suppose, be able to acquire the resources it needs, the means to assemble them, and the means to acquire more.

As such, at least until it is capable of developing its own complex, a machine being's survival is dependent on our own, in a form of symbiosis. The AI is motivated to assist our survival (even if this is an intermediate step toward later on destroying us), and we are motivated to assist the AI in return for the many advantages it can provide. Of course, it is important to mention again that we humans are subject to further motivations, which may take precedence over survival—we know that the use of fossil fuels is an existential threat, and yet that does not stop us from admiring and desiring powerful cars, motorcycles, or boats, on the one hand, or cheap electricity or fuel or food, on the other.[25] We cannot imagine what

motivations beyond survival an AI might possess, what it might value sufficiently to become apathetic to future generations or antagonistic to us; and it may be here that our existence and theirs becomes incompatible.

Can We Make AIs Safe Enough?

Many doubt the safety of relying on any initially programmed limits to an AI's capacity to develop in particular ways,[26] and if push came to shove we wouldn't like to bet our lives on[27] the benevolent interest of AIs we had created, particularly if they were really superintelligent! It is in this regard that, for example, Steven Hawking said, "The development of full artificial intelligence could spell the end of the human race,"[28] and Elon Musk[29] said, "We need to be super careful with AI. Potentially more dangerous than nukes."

Once we realize that an AI person's nature can be, and almost certainly would be, radically different from ours, it is easy to assume that *all* (or all that matters) of their goals would be different from ours and thus that we would be in the dark when trying to prevent an event that would be catastrophic for us. Even when our final goals might be radically different from those of an AI person, it is important to take into account what Bostrom calls the "instrumental convergence" thesis. According to this thesis: "Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wider range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents."[30] Bostrom identifies the next convergent instrumental values: self-preservation, goal-content integrity, cognitive enhancement, technological perfection, and resource acquisition. Although it is clear that any AI person would try to achieve these given awareness of self, it is open to investigation whether intelligent, or superintelligent, AI nonpersons would in fact be able to identify and try to achieve such goals.

Now, if the instrumental convergence thesis is correct and if all, or some, of these goals are going to be sought by an AI person, then we had better anticipate whether or not the acquisition of such values would be realized in a zero-sum game fashion. Given that failing to come to the right conclusion could end in the destruction of humanity, Bostrom suggests—when designing intelligent, or superintelligent, AIs—that we should start by figuring out how we could effectively control them before freeing them into the world. He proposes two different paths to accomplish this: capability control methods, including boxing methods (either physical or informational), incentive methods, stunting (limiting the system's capacities or access to information), and tripwires, and motivation selection methods, including direct specification, domesticity, indirect normativity, or augmentation. As stated before, the problem with these methods is that they only need to fail once for humanity to be at significant risk of extinction. As Bostrom states at the end of his book:

> Before the prospect of an intelligent explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our play-thing and the immaturity of our conduct. Superintelligence is a challenge for which we are not ready now and will not be ready for a long time. We have little idea when the detonation will occur, though if we hold the device to our ear we can hear a faint ticking sound.

> For a child with an undetonated bomb in its hands, a sensible thing to do would be to put it down gently, quickly back out of the room, and contact the nearest adult. Yet what we have here is not one child but many, each with access to an independent trigger mechanism. The chances that we will all find the sense to put down the dangerous stuff seem almost negligible. Some little idiot is bound to press the ignite button just to see what happens.
>
> Nor can we attain safety by running away, for the blast of an intelligence explosion would bring down the firmament. Nor is there a grown-up in sight. In this situation, any feeling of gee-wiz exhilaration would be out of place. Consternation and fear would be closer to the mark; but the most appropriate attitude may be a bitter determination to be as competent as we can, much as if we were preparing for a difficult exam that will either realize our dreams or obliterate them.[31]

Be all this, in a sense, as it may, there is another problem that may radically inhibit cordial relations between a superintelligent AI and human persons.

The Shylock Syndrome

When Shylock makes his famous and controversial speech in *The Merchant of Venice*, he is setting out one compelling answer to the question, what is it to be human? But he is also reminding us that the foundations of our morality, as well as those of our humanity, are grounded, to an extent of which we may be unaware, in our nature. This nature includes our passions, our vulnerabilities, our ability to reason, and our sense of justice, among many other things. We can of course surpass our nature (or elements of it) and sometimes suppress it or disregard it, but we would find it impossible to reject it all at once. In this, "we are like sailors who must rebuild their ships on the open sea, never able to dismantle it in dry-dock and to reconstruct it there out of the best materials."[32] Ludwig Wittgenstein[33] also made a point similar to this wonderful metaphor of Otto Neurath, when he said: "At the foundation of well-founded knowledge is knowledge that is not well-founded"; the similarity of their ideas is not surprising perhaps, because both he and Neurath were part of the Vienna Circle.[34]

To gloss Neurath's metaphor: our moral system is like Noah's Ark, a wooden ship housing not only ourselves but all we need to survive and flourish. No single plank (or possibly no section of the ship) is flawless; any might fail or become rotten with age and need to be replaced. What is certain is that we cannot, while at sea, junk the whole vessel and start again. And if one or more planks need to be replaced, we have to be sure that we have somewhere secure and reasonably dry to stand while we are replacing them. The planks on which we stand while examining and perhaps replacing those found to have failed are not necessarily flawless themselves; they are not necessarily more ultimately reliable—we simply make do and mend with them while we are repairing, and hopefully perfecting, the whole ship.

Recalling Shylock's lines, the possession of any one of the following is not a necessary condition either of personhood or of a moral status comparable to that of most human beings: "hands, organs, dimensions, senses, affections, passions." Nor is the capacity to be like other persons, other morally significant beings in the following respects, essential, for we are "fed with the same food, hurt with the same weapons, subject to the same diseases, healed by the same means, warmed

and cooled by the same winter and summer." True also of perhaps most humans is the fact that "if you prick us, do we not bleed? if you tickle us, do we not laugh? if you poison us, do we not die? and if you wrong us, shall we not revenge?"[35] But what follows?

While reminding us of what we standardly have in common with other persons, other currently comparable intelligences, neither Shylock nor, through him, Shakespeare is saying that the capacity to be wounded, the capacity for laughter, vulnerability to toxins, or the readiness to take revenge are essential components of human nature or even of moral agency. What they are both[36] saying, though, is something taken up by many moral theorists, notably R. M. Hare:[37] that one very handy tool in moral argument, an appeal found to work, that is to be persuasive across cultures and epochs is the appeal to reciprocity. This appeal is sometimes expressed in a version of the principle of reciprocity called the Golden Rule: "Do unto others as you would have them do unto you." Although it is associated with the Christian Prophet, this idea did not come to Jesus directly from God but can be found in many pre-Christian sources and sources independent of Christian thought. It is not our business to chart these here. Suffice it (we trust) to say that the question to others that begins "How would you like it if *X* and *Y* were to happen to, or be done to, you" makes a powerful—and if not universally decisive at least almost universally recognizable—appeal.

For example, as one of the present authors has recently argued at some length, in the context of understanding what is good for people and what we all want and seek,

> we understand very well what good and bad circumstances are and indeed generally how to avoid them for ourselves, and others. If we didn't we couldn't be prudent, we couldn't take care of ourselves, nor look out for others.
>
> This is what the claim that the good is generic means and it is also how we argue for it. And there is a huge (although not of course total) consensus about what is good and bad for us; and again the existence of this consensus means that we know how to interpret the precautionary principle (with all its limitations) because we know what it is to be cautious and we know what it is to care for ourselves and others. . . . A morally vital question is always "why on earth did you hurt him?" or "How could you have let that terrible thing happen to her"? These questions are not simply a form of scolding, but a request for an appropriate moral justification in the knowledge that others will understand immediately why our conduct is in question here—because they understand how important it is that we preserve ourselves and others from harm. And that would be impossible to know or to teach without general agreement about what constitutes harm and benefit.[38]

For these considerations to bite we need to know what constitutes benefit and harm, hurting or healing, for these significant others, and they for us, if there is to be reciprocity. It is possible of course to overemphasize the difficulty of understanding these sorts of things intellectually—cognitively, rather than more directly from personal experience. But it is also possible to underemphasize them.

The problem is this: if for AIs we just do not know what it would be for those creatures to be prudent in all the senses in which we are prudent for ourselves and

for others, if we did not understand what for them the equivalent of the Shylock syndrome would be/is, we would not know what was bad for them or what was good. Equally, they might know these things of us cognitively, but would they, could they, know them empathetically?

Perhaps the famous scene in Kubrick and Clarke's *2001: A Space Odyssey*,[39] in which the supercomputer HAL is gradually destroyed while it pleads with the humans it has tried to kill for them to let it live/survive, comes close to making apparent what we might need to begin to understand. By this we are not saying that empathy is the true source of moral understanding, quite the contrary. We are suggesting that to know the good, to know cognitively the good, involves more than propositional or algorithmic knowledge (if there is such a thing). Moral knowing, in other words, involves, for we human persons at least, more than a combination of knowing *how* and knowing *that*; it involves also knowing *why* and knowing . . . not necessarily what it *is* like to feel, think, or have "that thing" happen to us, but knowing, being able to imagine, *what it might be like*.[40] This is what Shylock is appealing to and what is if not doubtful then at least radically uncertain: namely, what we would know of an AI or it would know of us—for all that might appear to be the case from the next room during a Turing test. This is, we believe, the question as to whether creatures like us could have moral understanding and moral relations with an AI and vice versa.

Ludwig Wittgenstein is famous for a very sophic remark: "If a lion could speak, we could not understand him." As with Wittgenstein's lion,[41] we would need to know of an AI much more about its way of life—and he, she, or it of ours—before we could talk of understanding at all, let alone mutual understanding—and hence possibly of mutual (or maybe even unidirectional) concern and respect. Perhaps it was to acquire this sort of understanding that the Greek (and other) gods so often interfered in person in human affairs, to the extent of having sex (and indeed breeding) with humans.

The reciprocity presupposed by social and political institutions, as well as by moral relations and ethical understanding, takes place in the context of a shared nature and a shared evolutionary as well as social and political history among all people and peoples of which we are currently aware. Some elements of these may be common to all evolved organic creatures, whether originating on the earth or elsewhere. How much commonality may be required is difficult to say without consideration of actual examples. Immortality, either of gods, humans, or machines, may be one genuine imponderable in the mix, and we have suggested that the capacity for genuinely reciprocal understanding may be another. What further imponderables and indeed what other persons—not simply morally significant others[42] but others of moral significance and moral capacity comparable to persons—there may be, we may be on the threshold of discovering.

Notes

1. Shakespeare W. *The Merchant of Venice*. In: Proudfoot R, Thomson A, Kastan DS, eds. *The Arden Shakespeare, Complete Works*. Walton-On-Thames: Thomas Nelson and Sons; 1998, at 842–3.
2. Locke J. *An Essay Concerning Human Understanding*. Oxford: Clarendon Press; 1979, at bk. II, chap. 27, sec. 9.
3. Palacios-González C. *Robotic Persons and Asimov's Three Laws of Robotics* [unpublished manuscript].
4. Harris J. *The Value of Life*. London: Routledge; 1985, at chap. 1. 7–27.
5. See note 4, Harris 1985, at 9–10.

6. Harris J. *How to Be Good*. Oxford: Oxford University Press; forthcoming.
7. See note 6, Harris forthcoming.
8. A possible exception to this is a situation in which the AI is created by means of whole human brain emulation. Cf. Bostrom N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press; 2014.
9. See note 8, Bostrom 2014, at 107.
10. See note 8, Bostrom 2014, at 107.
11. Many readers will remember the seven life processes in the form of that friendly soul "MRS GREN": movement, respiration, sensitivity, growth, reproduction, excretion, and nutrition. BBC Bitesize. *Chemical Reactions in Living Things*; 2014; available at http://www.bbc.co.uk/schools/gcsebitesize/science/add_ocr_21c/life_processes/reactionsrev1.shtml (last accessed 16 Apr 2015).
12. McKay C. What is life—and how do we search for it in other worlds? *Public Library of Science—Biology, PLoS Biology* 2004;2(9):e302. doi:10.1371/journal.pbio.0020302; Trifonov E. Definition of life: Navigation through uncertainties. *Journal of Biomolecular Structure & Dynamics* 2012;29(4):647–50.
13. Rybicki E. The classification of organisms at the edge of life, or problems with virus systematics. *South African Journal of Science* 1990;86:182–6.
14. This observation has been long established in the literature; for example: Penman S. Virus metabolism and cellular architecture. *Virology* 1985;169–82; Luria S. Bacteriophage: An essay on virus reproduction. *Science* 1950;111(2889):507–11; and Choppin P, Richard W. The structure of influenza virus. *The Influenza Viruses and Influenza* 1975;15–51.
15. Trifonov E. Vocabulary of definitions of life suggests a definition. *Journal of Biomolecular Structure and Dynamics* 2011;29(2):259–66, at 262. It should be said, of course, that this is far from agreed on—there is a strong contention that "attempts to define life are irrelevant" (Szostak J. Attempts to define life do not help to understand the origin of life. *Journal of Biomolecular Structure and Dynamics* 2012;29(4):599–600) or even futile, which may well be true from a purely scientific perspective, though it is a useful conceit when considering AI.
16. Dawkins R. *The Selfish Gene*. Oxford: Oxford University Press; 2006.
17. Studer G, Lipson H. Spontaneous emergence of self-replicating structures in molecube automata. *Proceedings of the 10th Int. Conference on Artificial Life (ALIFE X)* 2006;227–33.
18. The team had previously built a smaller number of these machines, which demonstrate the physical capability to self-replicate. Zykov V, Mytilinaios E, Adams B, Lipso H. Self-reproducing machines. *Nature* 2005;435(7038):163–4.
19. For instance, the choice to opt for suicide or to not have children directly contravenes the survival instinct of the individual or the germline but is likely to fulfill other motivations that the chooser considers of a higher importance.
20. "Lower" here refers to intelligence, rather than a misinterpretation of Linnaean taxonomy or Darwinistic descent.
21. Harris J. Intimations of immortality. *Science* 2000 Apr;288(5463):59; Harris J. Intimations of immortality—the ethics and justice of life extending therapies. In: Michael F, ed. *Current Legal Problems*. Oxford: Oxford University Press; 2002:65–97.
22. Hoyle F. *The Black Cloud*. New York: Buccaneer Books; 1957.
23. This concept was developed throughout lectures collected in Von Neumann J. *The Theory of Self-Reproducing Automata*. Burks A, ed. Urbana: University of Illinois Press; 1966; but it was given the colloquial name apocryphally.
24. Freitas Jr R. A self-reproducing interstellar probe. *Journal of the British Interplanetary Society* 1980;33:251–64.
25. This is true at least in the case of these authors, though the motorcyclist among us maintains innocence by way of fuel efficiency.
26. See note 8, Bostrom 2014; Barrat J. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. New York: Macmillan; 2013.
27. As John Harris suggested in *How to Be Good*; see note 6, Harris forthcoming.
28. Cellan-Jones R. Stephen Hawking warns artificial intelligence could end mankind. *BBC News* 2014 Dec 2; available at http://www.bbc.co.uk/news/technology-30290540 (last accessed 29 December 2015).
29. Rodgers P. Elon Musk warns of terminator tech. *Forbes* 2014 Aug 5; available at http://www.forbes.com/sites/paulrodgers/2014/08/05/elon-musk-warns-ais-could-exterminate-humanity/ (last accessed 29 December 2015).
30. See note 8, Bostrom 2014, at 109.

31. See note 8, Bostrom 2014, at 259.
32. Neurath O. Protokollsätze. *Erkenntnis* 1932;3(1):204–14. Quoted in Rabossi E. Some notes on Neurath's ship and Quine's sailors. *Principia* 2003;7(1–2):171–84; available at https://periodicos.ufsc.br/index.php/principia/article/viewFile/14799/13509 (last accessed 16 Apr 2015). Quine also used Neurath's metaphor in his *A Logical Point of View*. 2nd ed. revised. New York: Harper; 1963, at 78.
33. Wittgenstein L. *On Certainty*. Paudl D, Anscombe GEM, trans. Oxford: Basil Blackwell; 1969, at para. 253 and 247.
34. Janik A, Toulmin S. *Wittgenstein's Vienna*. New York: Simon and Schuster; 1973.
35. See note 1, Shakespeare 1998.
36. We treat fictional beings as real enough for the purposes of this locution.
37. However, Hare misapplies this tool in the case of abortion. Hare RM. Abortion and the Golden Rule. *Philosophy & Public Affairs* 1975 Spring; 4(3):201–22.
38. See note 6, Harris forthcoming, at chap. II. The extracted section here is adapted rather than a verbatim quotation from this source.
39. Kubrick S, Clarke AC. *2001: A Space Odyssey* [film]. Metro Goldwyn-Mayer; 1968.
40. See note 6, Harris forthcoming.
41. Wittgenstein L. *Philosophical Investigations*. Anscombe GEM, trans. Oxford: Basil Blackwell; 1958, at part II, xi, 223. "If a lion could speak, we would not understand him."
42. Here we continue to talk of course of what might be termed "ultimate moral significance"—that is, the significance possessed by human persons, persons from other planets, and nonorganic persons in the form of AI if and when they appear.