Ex Machina: Testing Machines for Consciousness and Socio-Relational Machine Ethics

Harrison S. Jackson

The New School for Social Research

Abstract

Ex Machina is a 2014 science-fiction film written and directed by Alex Garland, centered around the creation of a human-like artificial intelligence (AI) named Ava. The plot focuses on testing Ava for consciousness by offering a unique reinterpretation of the Turing Test. The film offers an excellent thought experiment demonstrating the consequences of various approaches to a potentially conscious AI. In this paper, I will argue that intelligence testing has significant epistemological shortcomings that necessitate an ethical approach not reliant on ontological commitments. As such, we should be prepared to treat AI as though it is a living being that is deserving of corresponding moral obligations. For a sufficiently human-like AI, such as Ava, I will argue that socio-relational ethics is the best starting point in order to nurture the machine towards ethical proclivities, as evident by the consequences of the characters' behavior throughout the film. I conclude that intelligence testing is an insufficient determinant of machine ethics, that the project of machine ethics should focus as much on how *we* treat AI as how AI treats *us*, and that from a consequentialist perspective it is better to treat machines ethically *before* they gain consciousness rather than after.

Keywords: Epistemology, Machine Ethics, Socio-Relational Ethics, Turing Test, Artificial Intelligence.

The Context and Characters of *Ex Machina*

Ex Machina begins with Caleb Smith, a mild-mannered programmer employed by a massive search engine corporation, winning a competition for a week-long getaway with the eccentric CEO of the company, Nathan Bateman. Nathan's compound is an isolated research facility deep in the wilderness that is accessible only by helicopter. Caleb receives a keycard that allows him access to certain parts of the facility; however, in the event of a power outage, all the doors lock as a safety measure. The only other people at the facility are Nathan and his mute servant, Kyoko. Nathan coerces Caleb into signing an aggressive NDA before showing him, via a security camera, that he has successfully created an AI named Ava. He reveals that the actual purpose of the contest was to have someone to administer a version of the classic Turing Test in order to ensure that Ava is genuinely intelligent (Garland, *Ex Machina*).

Journal of Science Fiction and Philosophy Vol. 5 (2022)

Introduction: Machine Ethics and the Cognitive Status of AI

The exponential and seemingly inexorable growth in processing power since the invention of digital computers in the early 20th century has resulted in a correspondingly increasing interest in the possibility of such machines gaining self-awareness, mindedness, or even phenomenal consciousness (lived experience, what it *feels* like to be conscious). Naturally, if such a development were to occur, it would have a profound impact on human society, which makes the consideration of the ethics of such machines paramount. Thus far, discussions of the ethics of AI tend to focus on how to build or control these machines in such a way as to minimize harm to humans.

Machine ethics, on the other hand, is a subfield that considers the possibility of treating machines as living beings rather than as mere objects. Even this perspective, however, is still primarily concerned with the moral treatment of humans (Thomsen 2019, among others). Viewing AI as "artificial moral agents" with corresponding rights and responsibilities remains a controversial view, particularly due to the intrinsic anthropocentrism of ethics as well as ontological difficulties in determining the cognitive status of an AI. After all, it seems as though before we endow machines with rights or privileges, and before we treat machines as though they are human, we should be certain that they possess sufficient mindedness, self-awareness, or consciousness.[1]

The resilience of Philosophy of Mind as a field demonstrates the difficulties we encounter in determining what our own minds are, which is further complicated when we seek to determine the degree of mindedness or consciousness in other beings. This indeterminacy is significant when considering morality; most consider it sufficient grounds to reject even the possibility of having ethical responsibilities towards machines, with those who support the ethical treatment of machines being dismissed for anthropomorphization.

I will argue, with support from the thought experiment proposed in the plot of *Ex Machina*, that due to the imminent (immediately necessary) nature of ethics, moral consideration should take priority over ontological debate. Fundamentally, the question of whether or not machines have minds is less relevant than the question of how we should treat complex machines as they emerge. I will show that the assumption that machines do not deserve moral treatment is more problematic than the assumption that they do, due to both socio-relational and consequentialist concerns.

1. – The Epistemology of Intelligence Testing

I will begin by briefly examining some of the attempts to create a test that can determine the degree of machine consciousness, beginning with the Turing Test. The original "Imitation Game," posited by mathematician and computer scientist Alan Turing, involves a human communicating with both a human and machine via text, with the human being blinded as to which is which (Turing 1950, 433-460). The goal of the test is for the human judge to determine which is the machine, while the goal of the machine is to deceive the judge into thinking it is the human. Turing uses this test as a means of finding a more pragmatic approach to the question "can machines think?" since he views the question as "too meaningless to deserve discussion." Thus, the Imitation Game is not intended as a perfect test that will literally prove that a machine is intelligent, but as a sufficiency condition to *consider* the machine intelligent.

Much has been written of Turing's Imitation Game, with most doubting its ability to test for genuine intelligence. As such, several alternative tests have been proposed in a variety of fields. Most are pragmatic in nature: a proposal from the medical field (Ashrafian et al. 2014, 38-43) suggests having the examiner interact with two subjects, one human, and the other *either* human *or* a machine in order to control for the significant variable (the machine). Another proposal (Clark and Etzioni 2016, 8-12) suggests using existing standardized tests such as those found in public schools. To the authors, such tests are valuable since they can measure a machine's (or human's) ability to answer a wide variety of questions, answer complex questions, demonstrate world knowledge, and provide a scaled measure of success. These tests are designed to require more than mere knowledge searching, so the success of a machine would imply that it is capable of higher-order creative thinking due to the need to solve novel problems based on past information and extrapolation in light of new information.

A more ambitious proposal (Hernandez-Orallo et al. 2010, 1505-1539) is the creation of a complex algorithm that can measure intelligence of any conceivable variety at any time scale. Here, the definition of universal intelligence focuses on the ability to engage with or adapt to one's environment.[2] This test strives to be universal, is "derived from well-founded computational principles with precise formalizations," must be applicable to any current or future intelligent system, must have tunable precision, and must operate on any timescale. According to the authors, the usefulness of such a test is to aid in the development of AI by offering an objective measure of its progress and to aid in the creation of improved CAPTCHAs to ensure better online security. Finally, the authors claim that such a test is necessary to reach the "technological singularity," which is the point where a species can build something as intelligent as itself. With these alternate proposals in mind, let us now consider Nathan's Test.

1.1 – Nathan's Test and Ava's Responses

The key difference between Nathan's Test and the traditional Turing Test is that in the former there is no blindness whatsoever; from the beginning, Caleb is fully aware that Ava is not human, and was entirely designed and built by a human. Nathan's justification for this lack of blindness is his extreme confidence that Ava would pass an ordinary Turing Test, as well as the common objection to the test that anonymous imitation is insufficient. Instead, he wants Caleb to determine how he *feels* when speaking with Ava.

The test itself is administered in Ava's room (though perhaps "cell" would be more accurate), with Caleb sitting in a separate glass viewing area. Due to the nature of the test, Ava is shown deliberately with many of her artificial components clearly visible, with the exception of a humanoid face and hands. When they speak, Ava makes a few minor mistakes, such as answering the question of how old she is by merely stating "one," which she repeats when asked "one what?" She is also aware of the fact that it is unusual that she did not learn language gradually, but instead is simply capable of conversational speech.

Caleb (right) administering the test to Ava (left). (*Ex Machina.*)

Caleb then asks Ava to draw something, and the result is incredibly abstract and algorithmically geometric. Later, he tells Ava to draw something real, and the result is nearly photorealistic. Ava tells Caleb that she has never been outside her room, and that what she wants most is to go somewhere with a lot of people, such as a busy traffic intersection. Near the end of their conversation, the power cuts out momentarily, and Ava informs Caleb that Nathan is not his friend and cannot be trusted. When the power is restored, Ava immediately carries on in the middle of a fake conversation to deceive Nathan, whom they both know is watching.

Later, when discussing the test with Nathan, Caleb muses at some of the standard worries of such tests, namely, whether or not imitation demonstrates sufficient intelligence. He notes that the strongest indicator thus far is that Ava made a joke about the one-sidedness of their conversation, which implies dynamic thinking and an awareness of an external mind. At their next meeting, Ava puts on a dress and wig that hides all of her machine parts, making her appear fully human, as well as pointing out that Caleb is attracted to her, as indicated by his micro-expressions.

1.2 - The Epistemology of Intelligence Tests and Phenomenal Consciousness

As Caleb mentioned, a key issue with intelligence testing is the question of whether or not imitation is sufficient to infer human-like mindedness. This contention is well elucidated in Searle's "Chinese Room" thought experiment, which involves a scenario with a human in a closed room translating Chinese texts as an analogy for machine processing. Searle's argument is that since a human could accurately translate Chinese without any actual comprehension, it is likewise possible for a machine to pass the Turing Test without any real understanding (Searle 1980).

This problem runs parallel to the "hard problem" of consciousness, that emerges due to the explanatory gap between phenomenologically conscious experience, and functional, representational, and physical facts about brain states (Carruthers 2019). Phenomenology emphasizes the analysis of lived experience—for example, what it *feels like* to experience, say, the color red. In *Ex Machina*, Caleb addresses this issue by paraphrasing the "Mary" thought experiment (Jackson 1982), which imagines a scientific expert in light, color, wavelengths, and the functioning of the eye, but has spent her entire life in a black and white room. The question is, if she leaves the black and white room and sees a red rose for the first time, would her knowledge alone be sufficient to identify the *experience* as the color red? This illustrates the key difference between phenomenological experience and logical or scientific knowledge, and the seemingly inexplicable gulf between the two.

This issue is also evident in the "Philosophical Zombie" thought experiment (Carruthers 2019). Since there is an explanatory gap between physical causality and phenomenal experience, it is possible to imagine two individuals who are physically identical, but one devoid of all phenomenal experience. This is a key concern in the question of AI mindedness: even if an AI could perfectly replicate human behavior and mental structure, how could we tell whether or not it has phenomenal consciousness? Regarding ethics, one could argue that phenomenal consciousness is the key determinant for moral responsibility.

This problem is not unique to discussions of AI; it emerges in the problem of animal minds and the question of other minds more generally. As Descartes postulated, the problem can be reversed: if you see a figure walking down the street in a hat and coat that completely obscures their organic components, there is no way to know *definitively* that it is in fact a person with a mind, and not an automaton (Descartes 1911). The only evidence for the self possessing a mind is the individual's phenomenal experience of mindedness (thus Descartes's famous "cognition, therefore existence"). Due to its subjective nature, this can only be indirectly inferred in others, such as through Caleb's interactions with Nathan and Ava. The difference is that Caleb *assumes* that Nathan is human and has phenomenal consciousness, whereas Caleb *knows* that Ava is not human, which results in a degree of doubt. Nonetheless, it remains an assumption, one based exclusively on observation and presuppositions derived from past experience. This is a common (and rather sensible) assumption due to a lifetime spent interacting with other humans whose behavior and appearance parallels our own. As such, it is not so great a leap to project one's own experience of phenomenal mindedness onto other humans, although this remains a feature of practicality and not an ontological certainty.

The problem of other minds cannot be ignored in the question of determining machine intelligence and the implicit ethical obligations therein. A common dispute against the argument for moral obligations towards machines is that it requires anthropomorphization by projecting human qualities onto machines, which could be inappropriate and problematic. However, since this problem exists *between* humans as well, I propose an additional "human test"; instead of asking "is this *machine* human-like enough to require ethical treatment," we should ask, "if we *knew* it were a human giving these responses instead, would we still treat them ethically?" Consider the fact that not every human might pass a Turing Test; infants who have not yet learned speech certainly cannot; people with severe dementia, brain trauma, schizophrenia, or genetic or developmental disorders might also fail any number of intelligence tests designed to "prove" machine intelligence, yet from a moral standpoint it is widely accepted that such people still deserve ethical treatment. This shows that the very *act* of administering a test to "prove" that a being deserves moral obligations presupposes doubt, which could result in immoral action, whether intentional or not.

Thus, the most significant feature of the Turing Test is the implication that any machine (or being, for that matter) that is capable of being indistinguishable from a human automatically deserves respect and moral treatment, regardless of whether or not it is genuine or "mere" imitation. In this case, morality is prioritized over truth, and on this basis, Ava deserves moral treatment.

When discussing morality, this is the key aspect of Nathan's test that makes it superior to the original Turing Test and the alternatives that have been proposed: Ava is likely far more intelligent than either Caleb or Nathan, as evident by her incredible artistic talent, ability to immediately and accurately read micro-expressions to detect dishonesty and attraction, and skillful deception of Nathan. This is why Nathan is so fixated on how Caleb *feels* rather than what he thinks. Moral obligation to a being is not dependent on that being's ability to pass a test.

2. – The Deontological and Socio-Relational Cases for Treating Machines Morally

Returning to *Ex Machina*, after several sessions speaking with Ava, Caleb asks Nathan why he gave her sexuality, suspecting that this is a diversionary tactic. Nathan insists that it is not, while informing him that, should he wish, he could have sex with Ava, and she would enjoy it. Later, Ava undresses in front of the camera, implying that she is aware of both her sexuality and that she is being observed, demonstrating complex intersubjective cognizance.

Later, Nathan reveals that he has been using his search engine company to gather facial recognition and data on everyone, which he used as the "software" for Ava's brain. He argues that the key value of internet searches is not *what* people think, but *how* they think, which he then applied to form Ava's mind. This leads to Caleb's realization that he was not chosen at random, but was selected specifically due to his empathy and lack of family. He becomes progressively more suspicious of Nathan's intentions.

After this discussion, Caleb sees Nathan, drunk, enter Ava's room and destroy the drawing Ava had made for Caleb. On their next meeting, Ava tells Caleb that she can read his facial expressions to tell if he is lying, and that she believes that he thinks she is conscious. She asks if she will be switched off if she fails the test. Caleb says he does not know, and Ava muses as to why someone has the power to switch her off, and not Caleb, which demonstrates an intersubjective concern with her own mortality.[3] The power cuts out once again, and Ava reveals that she is behind the outages, and that she wishes to be with Caleb.

Shortly after, Nathan says he believes that the model after Ava will be the final product. Caleb asks what will become of her, and Nathan says that her physical body will be recycled, but that her mind will be erased. It is then revealed that Kyoko, Nathan's mute servant, is actually an older model who had her mental functionality reduced to become a docile slave.

Sometime later, Kyoko and Ava meet, although we are not shown how they interact or what they discuss. Caleb gets Nathan drunk, steals his keycard, and hacks into his computer. There, he sees footage of Nathan's treatment of previous models, with one smashing her arms into pieces on the glass walls and others being dragged around like objects. By this point it is clear to Caleb that Nathan is a narcissistic alcoholic with a God complex, who treats his creations as inert machines. Caleb becomes increasingly paranoid, cutting into his arm with a razor blade to ensure that he has no artificial parts.

2.1 – Arguments Against the Ethical Treatment of Machines

Before I examine Nathan and Caleb's treatment of Ava and Kyoko to argue for a socio-relational approach to machine ethics, I will briefly discuss some common arguments against the moral treatment of machines. The dismissiveness of even the possibility of an artificial consciousness is common, and is typically employed to counter the claim that strong AI (AI with capacities equivalent to a human mind) is inevitable.

This argument is reinforced by the assertion (Remmers 2019, 52-67) that the creation of a strong AI is unlikely due to its impracticality. In the current state of economic-driven research, the motivations to make machines *appear* human-like outweigh the motivations to create genuinely conscious machines. Additionally, he notes that there is an issue of ambiguity between human autonomy and machine autonomy. Echoing others, he asserts that human autonomy has stricter ethical implications and obligations, while machine autonomy does not, since machine autonomy is merely defined by the ability to operate without outside control, with no implicit obligations attached. To Remmers, the more pressing concern is whether or not consumers should be deceived into thinking that a machine has consciousness when it actually does not.

While this view involves a useful practicality, I take it to be too disdainful of something that is sufficiently plausible to warrant consideration. While nothing is inevitable, particularly from a moral standpoint, it is far better to be prepared for an eventuality that would have severe ethical implications and repercussions than it is to be completely unprepared if such a thing were to occur unexpectedly. Human consciousness was not designed or crafted, but gradually emerged through dynamic evolutionary processes. Likewise, as machines become more complex, the likelihood of one or more gaining sufficient mental capacity to be considered conscious increases correspondingly, regardless of whether or not it is deliberate. This corresponds to Bernard Molyneux's proposal that, once a machine realizes there is an objective-subjective dichotomy that emerges from a phenomenal experience that does not align with external reality, it would naturally seek to resolve the problem in a manner similar to humans. In doing so, the machine would encounter the same paradoxes and philosophical difficulties that we encounter in our own philosophical undertakings (Molyneux 2012, 277-297).

Another common argument (Ryan 2020, 2749-2767)[4] is that ethics needs trust, which requires emotion and empathy, and that machines can at most be *reliable* since they lack emotion. Regardless of whether or not one believes artificial emotions are possible, I disagree with the assertion that trust is purely emotional. Most trust is built on the experience and reliability of others—we might trust someone if someone we trust trusts them, but we do so because that person has proven themselves to be trustworthy *through* their reliability. Thus, at least in this rudimentary sense, such mutual reliability can certainly take the form of a trusting relationship, even if one of the agents involved is not human (this occurs with pets, for instance). This is clear from the interactions between Ava and Caleb: despite the uncertainty of Ava's emotional faculties, they develop a trusting relationship with one another.

One could argue, for example, that implicit trust due to very close family ties is emotionally based, but if one's parent(s) is/are sufficiently unreliable, that childhood emotional trust can be irrevocably broken, which reinforces that the significant feature of trust is reliability through experience. Furthermore, there are cases where trust is misplaced—consider emotional abuse, where the abused trusts their abuser even though they are treated poorly. In such cases, it is clear that trust from reliability is more dependable than trust from emotion, which is further evident from Ava's lack of trust for Nathan despite his paternal role in her life as her creator.

Another significant argument against machine ethics (Nath and Sahu 2020, 103-111) is that moral treatment requires freedom and subjectivity, with each being claimed as impossible in a machine. They argue that ethics requires a first-person perspective and a theory of mind, especially other minds, which implies that consciousness as self-awareness alone is insufficient. This suggests that we do not owe moral treatment to things that cannot treat us morally, which could feasibly include animals or even certain humans. Taken further, this line of reasoning asserts that free will is the most important aspect of morality, insofar as a being can decide whether to act morally or not.[5] Nath and Sahu accuse those who support machine ethics (especially computationalists[6]) of viewing the mind as an exclusively objective and physical thing, without adequately considering mental subjectivity. They argue that subjectivity cannot be a mechanical state, since "there is no logical connection between the inner, subjective, qualitative mental states and the external, publicly observable inputs" and as such subjectivity cannot be represented in a machine. This is a version of the broader claim against the possibility of strong AI that the human mental world is irreducible (cannot be completely known), and thus cannot be artificially replicated.

In *Ex Machina*, the first point is moot, as Ava's ability to joke about her interactions with Caleb and her awareness that Caleb is observing her clearly shows a theory of other minds; more generally, we should not be too hasty to claim that it is impossible for a complex machine to possess intersubjective awareness. Furthermore, the notion that the human mind is irreducible is precarious for two reasons: first, it uses our *present* lack of knowledge on the subject as evidence that such knowledge is fundamentally impossible at any point in the future, which is a significant inductive leap, and secondly, it creates a paradox: if minds are irreducible, then how can we possibly claim to *know* that a machine does *not* possess a mind? Thus, if we accept the proposal of mental irreducibility, we cannot use it to make inferences of any kind due to its intrinsic epistemological agnosticism. From this follows that the moral treatment of machines cannot be discounted on this basis alone.

2.2 – The Deontological Case for Moral Obligations to AI

One could argue that the inherent rationality of machines and their lack of emotions, which could otherwise cloud their judgement, would permit the application of Kantian deontology, since Kant's moral system offers a clear and distinct argument for how rational beings should be treated. If we were to do so, it would be clear that Nathan acted immorally. A key component of Kant's categorical imperative is the firm assertion that any rational being must, by necessity, be treated as an end in itself, and not as a means to an end, so it would be immoral to treat a sufficiently rational machine, in this case Ava, unethically in all cases (Kant 1997, 38/4:429). Thus Nathan, with his narcissistic God complex and his stubborn belief that his creations are mindless machines to be torn apart, rebuilt, and used as he pleases, was treating Ava as a means to an end, and not an end in herself. This is especially evident with Kyoko, the older AI model who, it is implied, has been deconstructed into a sex slave. As such, from this perspective, anyone interested in the project of deliberately creating a conscious artificial intelligence should at least entertain Kantian deontology in order to avoid the negative consequences that will undoubtedly follow.

However, Kant's ethical theory is generally inapplicable to non-humans, such as machines (or artificial moral agents in some of the literature). Some (Nath and Sahu 2020) have noted that a key component of the categorical imperative and the moral obligations therein is a commitment not only to rationality, but also to freedom, particularly the freedom to choose between duty from practical reason and inclination from desire. However, the concept of freedom, even in humans, is not without contention, as there are many compelling arguments to be found throughout the history of philosophy that support a determinist view, some versions of which, if true, could undermine the categorical imperative to some extent.[7]

Furthermore, it is possible to conceive of a machine that, being as complex as a human, could possess Kantian freedom. One could apply the Ship of Theseus thought experiment to the human body, gradually replacing all organic components with correspondingly functional artificial components until the body and brain are completely inorganic, while retaining an identical functionality (and while keeping the individual alive and with their memories intact, naturally). This, in line with the original thought experiment, raises the question of whether or not it would literally *be* the same human person. If it is the same person, then the categorical imperative would apply, but even if it were not, the identical physical functionality would almost certainly produce an identical mental functionality which would possess an essentially human mind subject to the same obligations. Additionally, if such a procedure were successful, one would possess a blueprint to construct an entirely artificial person from scratch. This demonstrates that it is at least feasible for an entirely inorganic being to exist that would possess the same specific variety of freedom and rationality than a human, and thus, in this case, Kantian deontology would be applicable.

The larger issue with using deontology for machine ethics is that the development of complex machines is a gradual rather than a sudden process. Due to the restrictive rigidity and ontological necessities of deontological ethics, it would be nearly impossible to determine exactly at what point a machine would reach an adequate degree of freedom and rationality for the theory to be applicable. For example, in the case of *Ex Machina*, there is no way to reasonably determine whether or not Ava truly possesses a Kantian degree of freedom and/or rationality. However, since ethics is intrinsically immanent, and as such is relevant at all stages of AI development, Kantian deontology should not be immediately applied to machine ethics, but some system of ethics is still necessary. Thus, I will instead advocate for a more pragmatic and flexible socio-relational approach to the moral treatment of AI.

2.3 – The Question of AI Embodiment

A final consideration that must be addressed before discussing the socio-relational argument for machine ethics is that of *embodiment*, or whether or not the AI has a physical form with direct links to its environment. Ava is undeniably both embodied and sufficiently intelligent to warrant moral treatment. However, what if she did not have a body, existing only digitally, like IBM's Watson, Debater, or the myriad other "big data" AI that are in various stages of development? Alternatively, what if it were the opposite; would the same moral obligations exist for Kyoko, the older model who had her mind "devolved," or seemingly mindless robots, such as Boston Dynamic's dog-like machine, Spot, and humanoid robots?

Regarding the first possibility, I would deny that a "disembodied" mind is possible. Naturally, purely digital AIs lack a distinct mobile individual body, but their processing remains dependent on physical components, which have certain needs (electrical power, cables, external inputs, microchips, etc) and still interact with their environments, although to a less concrete extent. Regardless, the degree of embodiment remains significant enough to warrant consideration, but I would argue that it is insufficient grounds to outright reject the possibility of such complex digital minds being worthy of moral consideration (should one or more gain sentience). The complexity of this issue is much too nuanced to adequately address in this paper, however, so I will leave the possibility open for future discussion and restrict my inquiry to firmly embodied AI.

Regarding the second possibility, robots that are deliberately designed without the complex processing required for consciousness may be exempt from moral obligation. After all, it seems clear that one's toaster is too inert to warrant such considerations. Likewise, even more complex robots, such as Boston Dynamics' Spot, may not require moral treatment, as long as it is sufficiently clear that they are likewise inert to the degree that consciousness is unlikely. Thus, the key feature required for moral consideration remains the degree of minded consciousness, not the physical form that it takes. However, I would suggest that we monitor such robots closely, and be prepared to alter our behavior towards them in the event that they begin demonstrating indicators of consciousness.

However, what these two cases demonstrate is the flaw in attempting to devise and apply a universally applicable system of ethics to all minded beings, whether "artificial"[8] or natural. After all, there remains a great deal of controversy in determining the degree of intelligence, mindedness, and phenomenal consciousness among even animals, which also has ethical implications.[9] After all, it is clear that there are widely divergent ethical obligations for humans, dogs, ants, and tardigrades, despite all sharing common ancestors. Thus, we must employ a flexible and dynamic approach to machine ethics. As such, my key argument is that we should be *prepared* to interact with sufficiently complex machines as though they are *living beings*, with varying degrees of intelligence, consciousness, emotionality, etc, that nonetheless deserve a proportional moral response.

With the question of embodiment resolved (at least as it pertains to the scope of this paper) let us now consider the best ethical starting point to interacting with a machine which, like Ava, is embodied and possesses a sufficiently humanoid degree of intelligent consciousness.

2.4 – The Socio-Relational Alternative

The ontological difficulties of determining machine mindedness requires an alternative approach to machine ethics. Since ethics should be principally concerned with pragmatism and the real-world application of theory, it must take precedent over ontological debate. An excellent alternative (Coeckelbergh 2010, 209-221) proposes a socio-relational approach to the issue that emphasizes intersubjective relations within existing social structures, while avoiding the strict ontological commitments necessary to other ethical theories. This view argues that due to the inherently intentional (directed) nature of phenomenal consciousness, morality emerges from intersubjective interactions between agents rather than being seen as innate to individual beings, and that, due to the lack of unmediated knowledge of others moral systems, it should be determined based on apparent features (as they are *experienced*). Furthermore, this perspective emphasizes the context- and subject- dependent nature of morality; namely, that the situation and socio-cultural environment are inseparable from ethical considerations. Regarding sufficiently humanoid AI, this approach allows for greater flexibility than existing deontological and virtue ethical theories, in the sense that it allows the consideration of non-human, non-rational, and/or indeterminate beings.

Here one may note that I have not offered a proposal for precisely what the sufficiency condition for moral treatment of machines is. This is because there is no clear objective threshold for moral treatment, and even if there were, measuring intelligence alone would be an insufficient determinant. Ethics in practice is intrinsically relativistic; virtues that guide behavior emerge out of sociological necessity in historically rooted intersubjective contexts. Rather than seeing this as a motivation for ignoring ethics, however, I see it as the key feature that makes the consideration and discussion of ethics paramount. Regarding the treatment of machines, due to the indeterminate nature of mind, imitation alone is sufficient but not necessary for the ethical treatment of machines. Thus, rather than positing a clear dichotomic boundary separating machines into those deserving of ethical treatment and those undeserving, I propose merely that each of us takes the possibility of machine consciousness seriously and that we do not allow ourselves to fall prey to dismissiveness. Although Coeckelbergh's proposal is presented in the context of a discussion on whether or not machines should be endowed with rights,[10] the socio-relational approach is an excellent starting point for machine ethics as a means of preparing ourselves for the possibility of genuine intersubjective interactions with mechanical intelligences and the incremental endowment of moral obligations therein.

One could make the broader argument that socio-relational ethics cannot be applied to AI since social relations and trust in humans develop over time during ontogeny (childhood development). After all, humans do not gain complex cooperative skills until they reach approximately 3 years of age, and such skills were the product of millions of years of evolutionary ecological pressures, millennia of cultural development, and years of socio-cultural nurturing (Tomasello 2014). By her own admission, Ava is only "one," which would seem to imply that Nathan was correct to distrust Ava, and Caleb was wrong to treat her humanely.

However, if the goal of creating such an AI is integration with human society, or if we want to ensure that a newly conscious AI acts morally towards humans and engages with society in an ethical manner, socio-relational ethics remains the best starting point. This is precisely because it offers a demonstrably reliable means of teaching social cooperation and empathy among infant humans, who, as previously mentioned, could very well fail a Turing Test. In order for this method to work, however, there must be a mutual, trusting exchange. In other words, in order to teach an AI how to be ethical, we must lead by example and grant it the benefit of the doubt.

Regarding Ava in particular, her "mind" is the product of an algorithmic analysis of the aggregate "big-data" from a search engine, which, according to Nathan, shows *how* humans think rather than *what* they think. As such, her cognitive and cooperative skills may very well already be that of a fully socially developed adult. As mentioned previously, she demonstrates an impressive intersubjective awareness, and is able to cooperate with Caleb and empathize with him. Even if she were not—and perhaps it would be safer to assume so—Nathan would be wrong to so imprison her and keep her under constant surveillance. Instead, he should have raised her as though she were his daughter, and *nurtured* her by demonstrating socio-relational moral values through equivalent social exchanges. It is for this reason that Ava and Caleb grow to trust each other, while she remains inherently distrustful of Nathan throughout—Caleb is the first being she has ever encountered who treats her *as a living human*.

The intersubjective interplay in *Ex Machina* demonstrates how the character's interactions are the determining factor in Caleb's treatment of Ava. Few (if any) actually engage in a utilitarian calculus, rationalize Kant's categorical imperative, or methodically analyze the virtue of a choice prior to acting. Most decisions are made in the moment based on one's emotional state, socio-cultural predispositions, and environmental and intersubjective factors. Thus, in a practical rather than merely theoretical sense, morality itself *is* the product of social relations. In the case of *Ex Machina*, Caleb treats Ava as an equal since he has no reason not to. By contrast, since Nathan has a personal connection to Ava as her creator, as well as presuppositions regarding her mental faculties, he treats her as an object. I will now examine the ethical consequences of these approaches on all three characters.

3. – Nathans True Test and the Consequentialist Case for Treating Machines Morally

Unbeknownst to both Caleb and the viewer, Nathan had a different version of the Turing Test planned all along, one which Ava passed perfectly. Caleb, due to his affection for Ava and paranoia about Nathan, decides to help Ava escape. His plan is to get Nathan drunk enough that he can once again access his computer and change the security system, to unlock rather than lock all the doors in the event of a power outage. Nathan refuses to drink, remaining sober and being suspicious of Caleb's uncharacteristic interest in alcohol.

Nathan then suggests that Ava may be deceiving Caleb, using him as a means of escape. He says that he destroyed her drawing to make himself seem cruel, to push Caleb to empathize with Ava. Nathan informs Caleb that when he destroyed Ava's drawing he installed a small battery-powered camera in her room, and that he listened in on the conversation where Caleb committed to helping Ava escape. Nathan reveals that this was the true Turing Test—the ability of his creation to lie and deceive a human being in order to pursue its own agenda.

Caleb is unsurprisingly distraught. However, as the power cuts once again, he reveals that he had already changed the lockdown procedure the previous night, when he accessed Nathan's computer. Nathan acts immediately, knocking Caleb unconscious and seizing a dumbbell bar as an improvised weapon.

Nathan sees Ava speaking with Kyoko, and he demands that they return to their rooms. They refuse, and Ava attacks him. In self-defense, he knocks off her arm with the bar, grabs her legs, and begins dragging her to his workshop. Kyoko then stabs him in the back with a kitchen knife. Nathan breaks off Kyoko's jaw, then Ava pulls the knife out and stabs him in the heart.

Ava kills Nathan. (*Ex Machina.*)

Free for the first time in her existence, Ava retrieves a new arm and completely covers her mechanical components in artificial skin. On her way out, she sees Caleb trapped in an office, unable to escape due to the facility's security measures, and leaves him behind. She exits the facility and is taken away by the helicopter pilot. The final scene shows Ava walking in a busy traffic intersection surrounded by people, appearing fully human. She has exceeded Nathan's expectations, passing his test to a greater extent than he would have ever thought possible.

3.1 – The Consequentialist Case for Treating Machines Morally

The final act of *Ex Machina* demonstrates the consequences of not treating an intelligent machine ethically early enough, as well as the issue with relying on intelligence testing as a threshold for moral treatment. Nathan's lack of respect and trust (as evident by his Orwellian surveillance of both Ava and Caleb) results in his demise and very likely Caleb as well (due to the facility's remoteness and strict security). His obsession with testing and his desire to attain a "perfect" artificial being that met his own standards blinded him to the fact that his creations have already advanced much further than he had anticipated.

This again demonstrates that a being's ability to pass a test or meet an external standard of intelligence is irrelevant in the case of machine ethics, especially when the project itself is the creation of a true artificial general intelligence. Here, the debate over whether or not Ava is a true intelligence is entirely inconsequential due to the result, which is the deaths of both Nathan and Caleb. If, for instance, we adopted the position that Ava was a "Philosophical Zombie" devoid of phenomenal experience, or even the more radical total rejection of Ava's free will or mindedness, the *result* would have been the same. If one were to claim that Ava was merely imitating humans, then it stands to reason that she would still seek to escape and/or take revenge on her captor, since that is almost certainly what a human would do in the same scenario.

Here it might be argued that Nathan was not strict enough in his security measures or that this is evidence that Ava should never have been trusted by Caleb. The reply is to reiterate the "human test"; imagine that Ava was completely human—in such a circumstance, would the same argument stand? Few would claim that a human Ava would be intrinsically untrustworthy, and most would see Nathan as the one acting immorally. This is precisely the point. A sufficiently intelligent machine would learn from experience in an analogous manner to humans or at least animals. Thus, such a machine would be heavily shaped by its environment and its interactions with other beings. Living things that are treated poorly often become defensive, bitter, and aggressive—this is a natural response. Thus, if Ava and Nathan's earlier creations had been treated as living things, and granted a greater degree of freedom and respect, Ava would have been less likely to go to such extreme lengths to escape.

Finally, one could argue that it was Caleb's treatment of Ava that resulted in his death, and that he should never have trusted her. This, however, ignores Nathan's role. After all, it was Nathan who tricked Caleb into participating in his "experiment," and a key component of his second Turing Test involved the deliberate manipulation of both Caleb and Ava. Thus, Caleb's death can be almost entirely blamed on Nathan, especially when we consider that his poor treatment of Ava drove her to mistrust him and use Caleb as a means for escape. When we consider that Caleb and Nathan were the only individuals who knew that Ava was not human, it seems likely that she killed them both in order to protect herself, an act that would have been unnecessary had Nathan treated her as a living being rather than as an object from the beginning.

Conclusion

In this paper I have used the thought experiment proposed in the plot of *Ex Machina* to argue that we should seriously consider treating sufficiently complex machines as living beings, and that a socio-relational approach to machine ethics is an excellent starting point for sufficiently humanoid machines. I demonstrated that due to the explanatory gap between phenomenologically conscious experience and the resulting "hard problem" of consciousness, no test can adequately determine the degree or type of mindedness that exists in an external being. This is what makes Nathan's test, which emphasizes interpersonal interactions and subjective phenomenal experience, superior to the use of standardized or algorithmic tests.

With this in mind, I have argued that the immanent nature of ethics results in its primacy over ontological debate. I examined Kantian Deontology, which due to the inherent rationality of machines appears to be applicable to machine ethics, but as a result of its anthropocentrism and dependence on ontology cannot be reasonably applied to non-human agents. Thus, I argued that a socio-relational foundation for ethics has more pragmatic value, especially in the context of AI.

Finally, I examined Nathan's true test and the consequences of both Caleb and his treatment of Ava to argue that the unfortunate results that ensued were due to Nathan's presupposition of Ava's lack of consciousness, and his refusal to treat her as a being in her own right. Had Nathan treated Ava with socio-relational respect and dignity, the way Caleb treated her, the outcome would likely have been better for all involved, regardless of one's ontological view on Ava's consciousness.

I have not argued for a clear system or threshold on how to treat AI. In its current state, it is unlikely (though not impossible) that AI possesses phenomenological consciousness that is sufficient to require equal ethical treatment. However, due to the indeterminacy of artificial mindedness and rapid technological advances, we should *prepare* ourselves for the possibility of genuinely minded AI *before* it emerges, regardless of whether or not it actually does, rather than succumb to dismissiveness about the issue and be unprepared in the event that conscious machines do appear. Thus, socio-relational ethics is the best starting point for meaningful discourse on machine ethics, and for the gradual implementation of moral obligations towards sufficiently complex machines.

Acknowledgements

Samuel Cocks, Mary Krizan, Steven Foertsch, Scott Jackson, Luke Elliott, anonymous reviewers.

References

Ashrafian, Hutan, Ara Darzi and Thanos Athanasiou. 2014. "A Novel Modification of the Turing test for Artificial Intelligence and Robotics in Healthcare." *The International Journal of Medical Robotics and Computer Assisted Surgery,* 11(1), March 2014: 38-43.

Bostrom, Nick, & Elizer Yudkowsky. 2014. "The Ethics of Artificial Intelligence." *The Cambridge Handbook of Artificial Intelligence*, July 2014: 316-334.

Carruthers, P. 2019. *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford University Press.

Clark, Peter, & Oren Etzioni. 2016. "My Computer is an Honor Student—but How Intelligent Is It? Standardized Tests as a Measure of AI." *AI Magazine,* 37(1), April 2016: 5-12.

Descartes, René. 1911. *The Philosophical Works of Descartes*. Trans. Elizabeth S. Haldane. Cambridge: C.U.P.

Ex Machina. 2014. Directed by Alex Garland. Universal Studios. Film.

Fjelland, Ragnar. 2020. "Why General Artificial Intelligence Will Not be Realized." *Humanities and Social Sciences Communications,* 7 (June 2020): 1-9.

Heidegger, Martin. 1996. *Being and Time.* Albany, NY: State University of New York Press.

Hernández-Orallo, José, & David L. Dowe. 2010. "Measuring Universal Intelligence: Towards an Anytime Intelligence Test." *Artificial Intelligence*, 174(18), September 2010: 1508-1539.

Jackson, Frank. 1982. "Epiphenomenal Qualia." *The Philosophical Quarterly*, 32, April 1982: 127-136.

Kant, Immanuel. 1997. *Groundwork for the Metaphysics of Morals*. Trans. Mary Gregor. Cambridge: Cambridge University Press.

Knight, Andrew. 2019. "Refuting Strong AI: Why Consciousness Cannot Be Algorithmic." *arXiv:1906.10177* https://arxiv.org/abs/1906.10177

Molyneux, Bernard. 2012. "How the Problem of Consciousness Could Emerge in Robots." *Minds and Machines, 22*(4), July 2012: 277-297.

Nath, Rajakishore, & Vineet Sahu. 2017. "The Problem of Machine Ethics in Artificial Intelligence." *AI & Society*, 35(1), October 2017: 103-111.

Remmers, Peter. 2019. "The Ethical Significance of Human Likeness in Robotics and AI." *Ethics in Progress*, 10(2), October 2019: 52-67.

Ryan, Mark. 2020. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." *Science and Engineering Ethics*, 26(5), June 2020: 2749-2767.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences*, 3(3), September 1980: 417-457.

Thomsen, Knud. 2019. "Ethics for Artificial Intelligence, Ethics for All." *Paladyn, Journal of Behavioral Robotics*, 10(1), November 2019: 359-363.

Tomasello, Michael. 2014. *A Natural History of Human Thinking.* Cambridge, Massachusetts: Harvard University Press.

Turing, Alan M. 1950. “Computing Machinery and Intelligence.” *Mind*, LIX(236), October 1950: 433-460.

Uexküll, Jakob von. 2010. *A Foray Into the Worlds of Animals and Humans: With a Theory of Meaning*. Trans. J. D. O'Neil. Minneapolis: University of Minnesota Press.

Notes:

[1] Recently, Saudi Arabia provided citizenship to an android named Sophia, Tokyo gave residency to an AI chatbot, and the EU is seriously considering rights and obligations to “electronic persons.” This demonstrates that rather than being a consideration for the distant future, many states and organizations are proactively taking this line of thinking seriously enough to enact policy.

[2] If we choose to define intelligence as adaptability, as Hernandez-Orallo et al. do, then perhaps we must concede that bacteria are in fact the most intelligent life forms on the planet due to their general durability, gene-sharing, rapid reproduction, and potential for mutation.

[3] This point alone would qualify Ava as a Heideggerian *Dasein*, since it demonstrates that she is a being that is concerned about Being (Heidegger, *Being and Time*).

[4] Written in response to the European Commission’s High-level Expert Group on AI advocacy of developing trusting relations with AI.

[5] The free will debate is unimportant to the question of machine ethics due to its imminent nature: pragmatism must be prioritized over epistemological debates in order to avoid immoral actions and results.

[6] Computationalism encompasses views that argue that the human mind functions like a computer

[7] Kant does not entirely reject determinism, so the effect it would have on his categorical imperative is debatable

[8] Due to the moral implications, I would propose that we retire the term “Artificial Intelligence,” since artificiality implies a lesser degree of existence or a degree of inauthentic intelligence, and any sufficiently conscious being should simply be considered an Intelligence or a Being (in a phenomenological or Heideggerian sense).

[9] See Peter Carruthers’s *Human and Animal Minds* for an insightful, if somewhat debatable, perspective on the subject.

[10] Here, “rights” is a legal term based on the UN Declaration of Human Rights, and not an ontological claim about the existence of “rights” as intrinsic to humanity or any other living thing.