xCoAx 2020
Computation,
Communication,
Aesthetics & X

2020.xCoAx.org
Graz, Austria

Machine Patiency and the Ethical Treatment of Artificial Intelligence Entities

Philip Galanter
galanter@tamu.edu
Texas A&M University, College Station,
United States of America

Keywords: Artificial Intelligence, AI, Deep Learning, Ethics, Complexism, Moral Philosophy, Patiency.

In recent years there has been explosive growth in the realm of artificial intelligence or AI. With that has come a body of ethical concerns regarding human implications. However, this paper explores our actions towards AI systems. The concept of machine patiency is the notion that humans may have moral obligations towards AI systems as they become sentient. Five bodies of knowledge are inspected to set the landscape for future machine patiency research. These are (1) the history of human encounters with sentient others, (2) topics from the philosophy of mind, (3) topics from moral philosophy, (4) niche specialists who study AI and ethics, and (5) the nascent field of complexism. The paper closes with a provisional affirmation of machine patiency as plausible on the basis of both natural charity and rational non-contradiction.

1. Introduction

In recent years there has been explosive growth in the realm of artificial intelligence or AI. This has been primarily fueled by the invention of *deep learning* algorithms. These techniques are an extension of neural network technology, and neural networks date all the way back to the invention of perceptrons in 1958 by the psychologist Frank Rosenblatt. (Rosenblatt 1962) Neural networks operate by using a simplified model of biological neurons that are assembled into layers, and typically take sensory (or other) information as input, and yield classifications (or other analyses) as outputs. In between the input and output neurons are *hidden layers* of neurons that process information based on the connectivity and weights of the neuron network. Adjustment of those connections and weights is what allows a neural network to generate meaningful output. The process of that adjustment is called *training* or *learning*, and is generally done by exposing the network to real-world training data and making iterative adjustments.

Prior to about 2010 the general consensus was that neural networks with more than one or two hidden layers were impractical. (Smith 1993) This was because adding layers exponentially increased both the amount of data needed for training, and the compute time required to do so. However, improvements in compute power (especially using GPU technology), and the easy availability of data (thanks in part to the Internet revolution) changed everything. Suddenly those building neural network systems were free to add many more hidden layers (the depth in deep learning), and achieve feats of AI previously out of reach.

But deep learning-based AI has raised new issues regarding its impact on society. These include questions with moral entanglements such as:

- How do we manage the elimination of jobs, and fairly distribute the new wealth AI yields?
- Since the internal representations used by neural networks are inscrutable, how do we guard against impactful errors?
- How do we protect privacy given the massive storage and data gathering in AI?
- How do we maintain control over systems that may someday surpass our own intelligence?
- How do we prevent racism, sexism, and other forms of bigotry from being learned by AI systems?
- How can we protect ourselves from disinformation and “deep fake” forged media?
- How do we train machines to respond to ethically challenging scenarios like the trolley problem?
- How do we control the use of AI in warfare via autonomous robotic weapons?
- How do we defend our AI systems from attacks by hackers or international bad actors?

These are indeed vexing questions, but they overlook an entire class of other problems. They only include threats to humans, and ignore those to individual AI systems. The question raised here is the ethical treatment of AIs, and our giving them moral consideration similar to that we extend to other people. (*AI* can either refer to the field of artificial intelligence, or to specific AI systems in operation.)

Concern for the ethical treatment of AIs will be met by some with skepticism or even derision, and dismissed as an absurd question. One would expect statements such as:

- AIs have no awareness. Therefore, we don't have to worry about how we treat them.
- AIs have no free will. Therefore, they cannot participate in ethically-based relationships.
- Shouldn't we make sure all humans are getting moral consideration first?

One purpose of this paper is to go beyond these knee-jerk reactions, and to set the stage for giving the question serious consideration.

Most are familiar with the notion of ethical agency. It refers to the ability, typically assumed human, to take assertive action in transactions with others, and to do so within an ethical framework that yields due moral consideration. We typically expect adults to act in an ethical manner, or in other words, we consider adults to have moral agency. We typically do not, however, confer agency in the case of children. As a matter of upbringing we incrementally expose children to moral expectations, but because children are still in the process of developing their cognitive capacity, we withhold agency and other forms of autonomy for their own protection. Thus, we expect adults to exercise agency, but not children.

In moral philosophy, also known in this context as ethics, there is a less well known but related concept called *patiency*. A *patient* is simply a recipient who is due moral consideration in an (ethical) agent's decisions and actions. When one kicks a rock there is no patiency involved. The rock is not due moral consideration, and in this sense kicking a rock is neither right and good, or wrong and bad. But kicking a child would be an entirely different matter.

So only adults have agency, but both adults and children have patiency.

With the terms understood as above, we might expect self-driving cars to exercise a very limited form of moral agency. An example is a variation of the trolley problem. The scenario is something like this. As a high-tech car is self-navigating down the street, a child runs in front of it. The only option available to the car at the time is hitting the child or swerving off the road running over various adults. What should the car do?

Even most humans would find the moral calculus here murky, but nevertheless we would hope the car "does the right thing." By granting autonomy, we have conferred upon the car some small degree of agency. The car is expected to make a decision in a moral context. But is the car a patient? If someone intentionally smashed the windshield of the car, we might accuse them of many things. But most would not complain that the rights of the car itself have been violated.

If AIs approach human intelligence, or even beyond, will this remain the case?

1.1. About This Paper

In introductory texts on moral philosophy the problem of skepticism is typically addressed very early on. (Rachels and Rachels 2018, Blackburn 2003, Singer 1994, Shafer-Landau 2020) As religious conviction has receded from the intellectual landscape of the west, so too has a relatively simple yet solid foundation for ethics and morality. Any ethics discussion must face this realm of ambiguity and uncertainty. But in this paper the skepticism expressed isn't about our general ability to answer moral questions. The skepticism here is about the relevance, or even relative absurdity, of moral consideration extended to AIs.

A leading issue is that patiency, the right to moral consideration, seems bound to the factual question as to whether the candidate machine has awareness. Without awareness, often called *sentience*, a machine or other object cannot suffer, feel pleasure, or have any experience at all. Of course, even this standard is not a simple binary one. For example, murdering a human in their sleep would still be considered wrong. And the treatment of those in a coma or "brain dead" raises vexing problems in medical ethics. These are questions for another time.

So machine sentience would seem to logically precede considerations of machine patiency. Unfortunately, this leads to questions that are often the object of disdain because they are so absurdly intractable. In the philosophy of mind there is a (special) concept of "*zombies*." These are theoretical objects that behave as if they are sentient, even though they are in fact entirely without consciousness. If AIs approach *general artificial intelligence*, i.e. if they seem to broadly understand the unconstrained and ambiguous everyday world as we do, one might wonder whether or not they are zombies. Perhaps they act as if they are aware, but it's only a simulation of external behavior. Or perhaps they have the moment to moment sensation of awareness that we experience as sentience.

But note that logically one could just as easily ask whether other (apparent) humans have awareness. This could be referred to as the problem of *solipsism*, the notion that one's own mind is the only one that exists. Typically, solipsism is simply rejected by most people as being prima facia absurd. "Of course, other humans have awareness!" But when it comes to artificial intelligence many will take the opposite side. They will insist that considering AIs as anything other than zombies is prima facia absurd. "Of course, computers don’t have awareness!” Is this a kind of bio-chauvinism? Can we reason our way past this potentially arbitrary contradiction?

This paper does not develop a refutation of solipsism. In fact, there is no attempt to develop a final argument from first principles. So, there will be statements presented here as axioms or hypotheticals that might, in some other context, be inferences. The intent is to set the stage for future research, and to describe the current conceptual lay of the land for machine patiency. However, a position affirming machine patiency is presented in the final section as being at least plausible.

So why take on the issue of machine patiency at this time? Because it’s likely that the question will leave the realm of philosophical reverie, and enter everyday situations that involve practical real-world concerns. How long it will take to achieve general artificial intelligence is a matter of speculation. Perhaps it will take hundreds of years or longer. Or maybe it will appear much sooner. Here we are assuming it is just a matter of time, however vague the estimate. Once machine sentience seems upon us, life situations will force the question, and morally significant decisions will have to be made even where deciding to not decide is also a significant decision. There will be no ducking the issue of machine patiency in life.

If someday your AI expresses fear and terror and pleads to not be turned off, what will you do?

As an aside, it should be noted that this paper only takes a western approach. Because patiency touches on issues regarding consciousness, and eastern notions regarding consciousness are so different than those in the west, covering those contingencies will have to wait.

1.2. Is Ethical Consideration Towards Machines Even Plausible?

If the first step in exploring machine patiency is to nail down whether machines can be sentient, the cause is already lost. At this time, we simply don’t know how to verify sentience in others with certainty. Perhaps we will never know.

Some will raise Alan Turing’s “imitation game,” now commonly called the “Turing test,” in this regard. (Turing 1950) In what is now a famous article, Turing first asks “can machines think?” But he immediately retreats from that question due to the vagaries of defining the term “think.” (It’s interesting, but not obviously relevant here, that he also finds the term “machine” problematic.) To replace that question he operationalizes the issue by suggesting what we would now call a double-blinded test. A computer and a human are both hidden, but they can communicate with an interrogator via a teletype. The interrogator alternates asking each a series of questions without first knowing which is the computer, and which is the human. Then the interrogator gives an opinion as to which is which. If the interrogator can consistently identify the computer, then Turing suggests we say that that computer apparently cannot think. But if there is no apparent difference between the two allowing identification as to which is which, then Turing suggests we say that the computer can apparently think.

But notice that Turing has changed what we would normally mean by the word "thinking." His version of thinking doesn't, for example, require sentience. It merely requires giving responses as if sentience was involved. Turing essentially begs the consciousness question entirely, just as he said he would. As an aside, in the text he seems to think people would be more inclined to find the results of his test more compelling if the questions and responses were audible. But that is beside the point.

The philosopher John Searle's "Chinese room" thought experiment offers a similar objection. (Searle 1980) He imagines a room in which someone who doesn't understand Chinese accepts questions written in Chinese, and "answers" them by slavishly following a large set of English language instructions, and writing answers in Chinese without actually understanding the input or output. The parallel is clear. The Chinese room apparently converses in Chinese even though it lacks understanding, just as a machine in the hypothetical Turing test might seem to converse without actually thinking.

The truth is despite all manner of study by philosophers, computer scientists, neurologists, and others, we simply don't know at this time what the necessary and sufficient conditions are for sentience in humans, animals, or computers. Without such knowledge we can't discount the possibility of machine sentience. And if machine sentience may be possible, the ethical consideration of machines is activated as a legitimate concern.

2. Areas of Study Related to Machine Patiency

There are a number of sources of research and opinion that can contribute to the consideration of machine patiency. Here is a provisional list of some of those:

- We can inspect history for examples of encounters with other sentient entities.
- We can turn to philosophy of mind regarding sentience.
- We can turn to moral philosophy or ethics regarding human patiency.
- We can look for niche scholars specifically focused on AI and ethics.
- We can look to the nascent field of complexism for relevant systems and models.

It was noted that at best only a good first step is possible within this paper. In the following sections we will simply survey these related topics, note initial impressions regarding machine patiency, and organize the landscape for further research. And we will look for sign posts that point in the likely direction additional research can take. But we should take care to not mistake the sign post for the journey. The hypotheticals presented here will require significant additional effort in exploration. Nevertheless, as previously noted, machine patiency will be affirmed as plausible.

2.1. Patiency and Encounters with Sentient Others

We can ask history about our experiences with other entities with regard to how much or little, and when and where, patiency is granted. But we shouldn't assume these historical decisions are correct. An understanding of past mistakes (not) conferring patiency may prevent those same potential mistakes being made with AIs in the future.

Many human encounters with "the other" take place when cultures meet. Although it is arguably a relatively recent invention, the concept of race has acted to defamiliarize others, and can mark one group as outsiders relative to an in-group. And this, in turn, has provided cover for diminished patiency, and diminished patiency has allowed abominations such as slavery and segregation. (Smith 2015)

One of the most divisive conflicts in contemporary American society is the patiency granted, or not, to prenatal humans. That disagreement has obvious ties to the issue of abortion. It was previously noted here that children and adults are granted relatively equal patiency even though their statuses in terms of moral agency are significantly different. For many patiency doesn't begin at conception, and it only comes into play after many months of gestation if not birth.

There are also disputes on the other far end of lifespan. As people age and become infirm, society allows that they lose some agency, and rescinding that agency may appear to be a form of reducing patiency. (Mueller, Hook, and Fleming 2004)

Additional examples of encounters with sentient others include ongoing human relationships with the animal kingdom. Humans are capable of great kindness towards animals and especially domesticated animals. Sadly though, animals are more frequently subjected to what is objectively nothing better than torture and terror leading to premature death. Some feel this kind of treatment of animals should and will diminish and disappear over time. (Singer 2009, King 2017)

It's worth noting here that some find confidence in comparing human notions of suffering with subjective animal experience based on common underlying biological factors. This is especially relevant in terms of hormones, neurotransmitters, and brain organization where it seems that nature uses the same mechanisms over and over again. It is thought that subjective experience supervenes on brain activity. We may not understand the mechanics of consciousness, but for many having a common biological substrate increases the likelihood of a commonality in subjective experience. This would seem to position machine sentience at a disadvantage relative to animal sentience. This is related to, and somewhat denied by, the notion of *functionalism* discussed here later.

Finally, there are numerous depictions of encounters with non-human sentience in literature. Most obviously notable is the genre of science fiction.

But mythology and folk tales are replete with enchanted animals or golem-like creations that exhibit sentience beyond our everyday experience. While not based on historical facts, literature can reflect the concepts and ideas a culture is primed for. It may be the broader culture is only waiting for our computing technology to catch up.

(Some might think here of the novel *I, Robot* which introduced Isaac Asimov's "Three Laws of Robotics." (Asimov 1963) These, however, are only tangentially related to the issue of machine patiency because they specify obligations a robot is to fulfill, not obligations humans have towards robots.)

2.2. Philosophy of Mind and Ethical Patiency

Western philosophy also offers a variety of views regarding consciousness, sentience, and free will. The issue of machine patiency activates an already uncomfortable tension in contemporary society. On the one hand science describes a universe that, short of quantum effects, is mechanistic and subject to deterministic cause and effect. And western culture is nothing if not under the sway of science.

On the other hand, typical western moral values assume that ethical choices are indeed freely chosen. The conflict is sharp when considering the moral implications. For example, is it just to punish someone for apparently instigating an illegal event when all along that event was as inevitable as an apple falling from a tree?

For the time being here are some snapshots of some of the theories of consciousness, and as is relevant to this paper, tentative implications for machine patiency. It's important to remember that to date none of these theories provides a complete answer as to how conscious machines can be constructed, or how human consciousness works, or how third parties can verify a given entity has consciousness.

Quantum Emergence (Penrose) – Roger Penrose and Stuart Hameroff have suggested that biological structures are able to harness quantum phenomena to leverage superposition. (Penrose 2016) This allows the brain to essentially compute solutions to problems that are not amenable to Turing machine computation, and are limited by Gödel incompleteness. It is said that this yields consciousness. If true, we would know that a necessary but not sufficient condition for machine patiency is a capacity for quantum computing.

Panpsychism (Whitehead) – Panpsychism is the theory that consciousness is a fundamental, low-level, aspect of nature distributed throughout all matter or being. (Goff, Seager, and Allen-Hermanson 2017) From this point of view even subatomic particles have a trace of sentience that can be aggregated in larger systems. Alfred North Whitehead is the most notable modern advocate, and he integrated panpsychism as part of his process philosophy point of view. If true, machine patiency would probably have to be viewed as something awarded in a continuously varying amount.

Property Dualism (Chalmers) – Descartes famously wrestled with the mind-body problem which asks how it is that humans have both physical and mental aspects. He proposed what we now call substance dualism as a solution. Substance dualism proposes that mind and body are ontologically distinct substances. One problem with substance dualism is explaining how the two interact. David Chalmers has suggested that a single substance, the brain, can have two kinds of properties, mind and body. (Chalmers 1996) Unfortunately, it's hard to see how this paradigm provides leverage over identifying whether consciousness exists in a candidate for machine patiency.

Bio-Emergence (Searle) – John Searle endorses the notion that mind is an emergent property of the brain, and it creates a "first-person ontology" of sentience that is inaccessible to third parties. (Searle 1992) He contends that computing hardware may be able to simulate the operation of the brain, but it won't create the kind of first-person ontology we associate with consciousness. Thus "weak AI" which does not create sentience is possible, but "strong AI" which does is not possible. If true, one would think that machine patiency is a non-issue. However, due to the inaccessibility of first-person ontology, Searle would say there is no empirical way to verify this.

Functionalism (Putnam) – Functionalism is the contention that mental states are assembled purely on the basis of the functional relationships of its parts. Hillary Putnam was an early advocate for functionalism. (Putnam 1988) Mental states then are in a sense hardware agnostic, and minds constructed from biological neurology can also be constructed from computer electronics or any materials with the appropriate functional relationships. If true, candidates for machine patiency should qualify if properly designed, but it's not clear how an observer would know whether the constructed device was sentient or simply a zombie device.

Integrated Information Theory (Tononi) – Integrated Information Theory (IIT) presents a relatively new paradigm developed by starting with the phenomenon of conscious, and from that ultimately inferring what the physical substrate must be like to support it. (Tononi et al. 2016) One key to the theory is a measurement Φ (Phi) of the capacity for a system to integrate information. The theory is highly technical and controversial. It has traces of panpsychism in that tiny degrees of consciousness exist in small structures and can be distributed and of varying density. One significant criticism is that Φ measures an aspect that may be necessary, but is not obviously sufficient for consciousness. This is very much a work in progress, but it holds some hope that a means of measuring consciousness in a computer may become available. The relevent implication here is that IIT might provide an objective measure for assigning machine patiency.

Mysterianism (McGinn) – The so called "hard problem" of consciousness was identified by David Chalmers as the experience of what it's like to be something. It turns in part on the awareness of qualia, the redness of red, or the sweetness of sugar. The canonical example is captured in the title of Thomas Nagel's article "What is it like to be a bat?" (Nagel 1974) Some, like Colin McGinn, have concluded that consciousness, and particularly the qualia of the hard problem, are things simply beyond human comprehension. (McGinn 1991) If true it would not bode well for being able to identify machines worthy of patiency.

2.3. Ethical Patiency in Moral Philosophy and Ethics

Most philosophical systems require a commitment to rationality as being axiomatic, and that typically starts with the embrace of the three laws of logic. The first is the law of identity stating that a true proposition is (always) true. Next, the law of non-contradiction demands that if a proposition is true, it is not false. Finally, the law of the excluded middle insists that a proposition must either be true or false, and there is no third option. In some ethical systems (Kant 1950) the law of non-contradiction is specifically harnessed as a powerful tool.

Philosophy may not have much to say about specific irrational acts, but there is much that can be said in the form of meta-critique about irrationality itself. Individual irrational actors may claim to be driven by love, hate, or other emotions. Or they may claim to have been inspired by divine revelation, or some artistic muse. Some have offered that existentialism in philosophy offers a defense of a kind of irrationality. (Barrett 1990)

A commitment to ethical impartiality is often directly inferred from a commitment to rationality. By impartiality we mean that there must be a reason for differential treatment. However, there is typically a belief that ethics only applies to humans. The underlying assumption is that humans are uniquely rational. But this is exactly where AIs may challenge that default attitude.

Western philosophy offers various systems of ethics, and with each there is a theory, explicit or not, of patiency. These theories of patiency can precede, intertwine with, result from, or otherwise reflect the moral system they are associated with. A short inventory of moral approaches, similar to what might be expanded upon in introductory ethics texts, is presented here. It is offered not merely as a survey, but also as a way to add preliminary implications for machine patiency.

Nihilism, or more precisely *moral nihilism*, is the belief that there is simply no moral right or wrong. Although this would seem to be a dead-end, in ethics the defense against meaninglessness is a viable discussion. (Barrett 1990)) In any case, it may be that conferring patiency to AIs is in the practical self-interest of the human nihilist involved. A moral nihilist, for example, could enter into a social contract with an apparently sentient AI purely for reasons of self-interest. So, conferring machine patiency in that case might be possible, even if somewhat in bad faith. (And why should a nihilist worry about bad faith?) This approach can be related to social contract theory which is briefly discussed below.

Moral relativism, or simply *relativism,* is a bit like nihilism in that it denies that absolute normative judgements are available to us. But it carries with it the consideration that a given culture will have a strong commitment to a system of normative guidelines; and that while having no absolute basis, norms are nevertheless rigorously followed and enforced. Under such a system the issue of patiency, and thus machine patiency, could potentially take on any shape depending on cultural taste. (Gowans 2019)

Religion is potentially as variable as nihilism or relativism. Plato asks whether an act is moral because the gods say it is, or do the gods say an act is moral because they have the wisdom to see the truth of the matter? The second possibility raises the ethical good above the gods, and this is entirely compatible with Plato's metaphysics.

But in western monotheism, notably the three major Abrahamic religions of Judaism, Christianity, and Islam, there is no higher power than God. God is the ground of being, and God stitches morality into the fabric of being as part of the act of creation. Thus, right and wrong have ontological weight. Humans are viewed as uniquely possessing souls given by God, and because of that humans are conferred patiency by God.

Being without souls, animals cannot be said to have patiency. Humans have been given dominion over all living things by God. But that's not to say animal abuse can be justified. People are also charged with stewardship of the natural world. In addition, it is thought that cruelty to animals is likely to have a corrupting influence on a person's soul, and maintaining one's soul is of highest moral priority.

It would be terribly naive to suggest a single size fits all of western religion. But an initial signpost for sentient machines probably points in the same direction as that for animal patiency. As such patiency is probably not conferred upon apparently sentient machines, but cruelty towards sentient machines is to be minimized. (Detection of sentience, however, remains a problem.)

Kantian ethics is simply the study of ethics as authored by the philosopher Immanuel Kant. Kant is arguably the most important philosopher of the modern era, and his writing is the benchmark for *deontological ethics*. (Kant 1950, 1956) Deontological ethics is a rules-based practice. This is to be contrasted with *consequentialist ethics* which is a results-based practice. There are at least two propositions that serve as maxims central to Kantian ethics. First there is the notion that persons should never be treated merely as a means to an end, but rather as ends in themselves. This is related to what Kant sees as a uniquely human capacity for rationality. The other maxim discussed here is actually stated by Kant in a number of forms, and is referred to as the *categorical imperative*. Perhaps the most well-known formulation is this:

> “Act only according to that maxim whereby you can, at the same time, will that it should become a universal law.”

Kantian ethics is notoriously complicated and entire academic careers can be built on related research. Here the admittedly simplified view offered is that Kant asks whether, given a commitment to rationality, there is something about rationality itself that can give us guidance in our choice of moral actions. In this regard the categorical imperative gains its strength from the law of non-contradiction.

Kant afforded personhood to humans but not animals because he viewed humans as being uniquely rational, where animals are not. However, not unlike monotheistic religious views noted above, he viewed cruelty to animals as something to be discouraged because of its corrupting influence on the human committing the abuse.

This subtopic clearly invites additional work. But if an apparently sentient AI could be considered to be as rational as a human, then it’s hard to see why Kant wouldn’t afford it more moral consideration than an animal. This would mean that AIs would have to be treated as ends in themselves, and not treated merely as means to some human end. That would obviously be a radical shift in human attitudes towards computers. Quite possibly Kant would further view sentient AIs beyond some threshold of rationality as having both agency and patiency.

Utilitarian ethics is a consequentialist rather than deontological theory based on the notion that increasing happiness is the highest good. (Driver 2014) The operative phrase is often stated as “the greatest amount of good for the greatest number.” While many intuitively find this to be a strong and just principle, as well as a pragmatic guide that easily fits contemporary liberal society, there are a number of well recognized problems.

The established critique of utilitarianism is beyond the scope of this paper. However, from a systems point of view the formula “the greatest amount of good for the greatest number” is an underspecified multidimensional optimization problem. At best it offers the opportunity for a set of pareto-optimal solutions, but not a way to narrow those solutions to a single choice. In short, it does not provide a way to make final moral decisions.

Historically utilitarians have afforded animals consideration because they are seen as sentient creatures that experience pain and pleasure. Utilitarians tend to be especially sensitive to awarding patiency where suffering is likely. However, human needs are typically given priority. So, for example, one might speculate that utilitarians will not compel vegetarianism, but they will strongly prefer farm animals to be kept and slaughtered “humanely.”

Nevertheless, given their willingness to give animals some moral consideration, it seems reasonable to see utilitarianism as pointing in the direction of granting machine patiency.

Social contract theory advocates for an ethics of cooperation where individuals voluntarily enter into an agreement with others, and that agreement asks that they give up any anarchic disposition in exchange for the stability and safety of an ordered society. Ethical thinking in this context leaves the traditional metaphysical weight of "the good" behind, and instead inspects relationships and social networks in terms of the exchange of benefits. (D'Agostino, Gaus, and Thrasher 2019)

In an unordered society a person who purely cooperates with others will come out on the losing end because they will be predictable and ripe for exploitation. (This resonates with game theory as noted in a following section.) Humans as rational animals are therefor moved towards negotiation, and form social contracts where each individual is guaranteed a moderate benefit. Under such a contract, patiency is conferred by rational agents to each other out of enlightened self-interest.

Because a trust-based highly ordered society is also the most fertile ground for sporadic criminality, part of the social contract will tend to include a shared enforcement mechanism. It takes the form of a variation of the golden rule that is operationalized. "You will be done unto as you do unto others."

A known problem with social contract theory is that those who pose little threat, or have little to offer, may be left out of the contract and find themselves unprotected. This does not bode well for apparently sentient machines. The question becomes whether we will anticipate the need to include apparently sentient machines in the social contract, or whether such machines will have to "do (painful things) unto others" in order to demand inclusion in the social contract. Granting apparently sentient machines patiency in such a context seems underdetermined, but may be forced by enlightened self-interest just as it is in the human-to-human case.

2.4. Niche Scholars Focused on AI and Ethics

It's obviously important to include in these considerations any niche scholars specializing in AI and ethics. While the bulk of writing about AI and ethics focuses on human-to-human obligation, some is about the ethical treatment of AIs and machine patiency, including AIs embodied as robots.

Notable among these is Joanna J. Bryson. In the essay *Robots Should Be Slaves* she argues that AI-endowed robots would only suffer if they were programmed by humans to suffer, and they are best viewed as sophisticated tools for our use in a way that would be objectionable if they were humans. (Bryson 2010) In line with this notion of the ontological subservience of AIs, in a more recent article regarding machine patiency she concludes "We are therefore obliged not to build AI we are obliged to." (Bryson 2018)

In S*uperintelligence: Paths, Dangers, Strategies* (Bostrom 2014) the threat of runaway growth of autonomous artificial intelligence is considered. Bostrom has widely published on allied topics, and is one of the many experts who signed on to the Asilomar AI Principles agreement intended to head off all manner of AI related ethical and social problems.

Also worth mentioning is the work of ethicist Peter Singer. (Singer 2009) His work on animal rights parallels the same kinds of concerns we see in the consideration of patiency for AIs. It's entirely possible, if somewhat ironic, that the fresh view afforded by the consideration of machine patiency may reinvigorate the kind of concern for animals that activists like Singer have expressed for decades.

And indeed there are dozens of others actively publishing in this realm in philosophy, computer science, and other general disciplinary journals, as well as specialty journals such as the *Journal of Artificial Intelligence and Consciousness.* A definitive bibliography would be a greatly appreciated project in itself.

2.5. Complexism and Machine Patiency

Complexism is a nascent worldview that takes the findings and tone of complexity science, and uses same as a platform to analyze and critique the problem space of the humanities. (Galanter 2016) As such complexism has implications for ethics.

Previous work on complexism has yielded a new model of authorship that ventures beyond those from modernity or postmodernity's post-structuralism. (Galanter In Press 2020) This model illustrates when a generative art system would have to be credited as truly being an author. To the extent that the determination of authorship is "giving credit where credit is due," such an analysis is already on the precipice of conferring a degree of machine patiency. (There are, however, other formulations where identifying a machine as an author is purely descriptive and without moral connotations.)

Complexism views human relations through the lens of network analysis, and ethical behavior is viewed as an emergent product of co-evolution. Co-evolution emphasizes that when it comes to "survival of the fittest," properties that contribute to fitness are a moving target. As an environment changes, and as competition creates a kind of "arms race," each adaptation stimulates adaptation by others in response. Ethics in this light can be seen as a kind of network protocol in a co-evolving human network. As that network evolves, so too does the network protocol. Thus the "moral good" can change slowly over time.

But are there high-level protocols that are strongly persistent? An interesting question is the mystery of the Golden Rule, "do unto others as you would have them do unto you." Some form of the Golden Rule is found in virtually every society and religion. Why is that? One possibility is that the Golden Rule is an inevitable emergent property of group adaptation.

The canonical prisoner's dilemma problem from basic game theory can provide insight. (Kuhn 2019) It can be described as follows:

A situation occurs where criminal partners have been caught, and they are being interrogated separately under the following terms. If both remain silent, each will serve one year in prison. If each betrays the other, each will serve two years. But if one betrays and the other remains silent, the one who remains silent will serve three years, and the one who betrays the other will go free.

The paradox is that if both prisoners remain silent, the total time served will be minimized, but from a selfish point of view the optimal move is to betray. However, when an iterated version is played as a game, and players can remember previous rounds, what tends to emerge is a winning strategy called tit-for-tat. Initially player one will strategically give player two the benefit of the doubt, and will cooperate by remaining silent. If player two betrays, i.e. selfishly tries to profit at player one's expense, on the next round player one will betray player two. But if player two cooperates, then player one will continue to cooperate.

This is, in effect, an operationalized version of the Golden Rule; "I will do unto to others as they do unto me." The surface cooperative behavior may appear to be altruistic, but the underlying mechanism is one of enlightened self-interest.

It's worth noting there are also traces of the Golden Rule operationalized this way in observed animal behavior. (Schmelz et al. 2017, Choe et al. 2017, Warneken and Tomasello 2006) The extent to which this emergence is directly related to genetic inheritance is arguable, but it certainly would be advantageous to have such behavior hardwired rather than discovered anew by each individual.

Is it possible that what feels like empathy at the level of consciousness is in fact a somewhat inevitable result of this co-evolution? Empathy can be viewed as an adaptation promoting social bonding and cooperation yielding a survival advantage. But the individual's survival instinct also brackets empathy with limits that protect the individual from uselessly giving to the point of personal jeopardy. As a result, each human defines and defends a circle of empathy. Their capacity for empathy, i.e. the number of people within that circle, is speculated to be in part determined by their genetic inheritance. Those outside the circle become "the other", and are eyed with greater suspicion. At its edge this circle becomes the functional method used to define and confer patiency.

Some moral values must be as universal as breathing for the survival of any culture. For example, the importance of the care of children is a universal human value. Similar cross-cultural values, e.g. the prohibition of murder, lying, etc., are enforced within the circle of empathy. Evolutionary pressure ensures these protections will be enforced by the culture, but the most important and permanent values are suggested to be genetically inherited. (More precisely, simple neurological precursors are genetically inherited, and these lay the groundwork for the emergence of empathy and cross-cultural values. Sometimes something in that chain of causation goes wrong, and sociopathy follows.)

In a way similar to the iterated prisoner's dilemma, a trust-based highly ethical society creates fertile ground for criminality, because so many members of that society are available as unsuspecting cooperative victims. Even if a society somehow reaches one hundred percent compliance with the law, some intelligent agents will later discover non-cooperation to be profitable and put it into play.

An innate predisposition for an emergent circle of empathy is also hypothetically a biological basis for tribalism. In a relatively short amount of time homo sapiens has gone from small group living to coping with "the global village." Evolutionary change takes much longer, and the expansion of the optimal size for an inherited circle of empathy likely has not kept up. In addition, due to the size of our current society, many transactions are one-offs. This results in the iterated prisoner's dilemma model being undercut for lack of iteration. I.e. it's easier to cheat someone you will never see again than to cheat someone you expect to see repeatedly.

Finally, in every culture morality is not permanently fixed, and there is a slow ebb and flow of public ethics in creative tension. An area of investigation, perhaps via simulation, for both complexity science and complexism is whether feedback loops within a society lead to ethics as a chaotic system. Chaotic systems are deterministic (i.e. follow cause and effect) and yet are unpredictable because of sensitivity to initial conditions. The implication is that at any given time we can describe the general contours within which ethical behavior operates (i.e. the phase space), but individual moral decisions remain somewhat unpredictable to observers. Perhaps such social turbulence can be quantified and empirically verified.

So, what are the implications within the framework of complexism for machine patiency? Much of the above is informed speculation, but speculation nevertheless. Additional work establishing theory of mind and empathy in other animals will be crucial, as will be additional understanding as to how genetic information ensures this emergent behavior. A scientific basis for something akin to "human nature" is essential, and flies in the face of social constructionist attitudes in the humanities.

But with additional evidence, this line of thinking leads to the following idea. To the extent that people and AIs are in transactional relationships, and AIs have achieved what appears to be autonomy, rationality, and sentience, enlightened self-interest will suggest that humans extend moral consideration to AIs. Co-evolution within a circle of empathy is proposed to be part of our genetic inheritance. It has emerged because it tends to ensure optimized benefits for all, and avoids conflicts with "the other" that only ensure strife.

3. Going Forward: Rationality and Charity

The preceding sections have surveyed various theories in ethics and philosophy of mind in the context of machine patiency. Both fields currently include viable, yet conflicting, theories. In addition, complexism has been offered as a possible framework for integration and further research. That completes the initial goal of describing the landscape for future machine patiency research.

In this final section we will consider whether some provisional position regarding machine patiency can be suggested. Speculation aside, we are locked out and cannot access the possible first-person experience of AIs. This is not specific to AIs. It is the very nature of first-person ontology that third parties cannot know what it's like to be a bat, or an AI, or another human. But coming over the horizon is the likelihood that as a practical matter, we will face decisions relative to the treatment of AIs. Even deciding to not decide is loaded with the risk of irresponsibility.

In every culture other humans are assumed to be sentient and to have awareness that is similar to our own. This is referred to in psychology as a "theory of mind". Theory of mind emerges in children in the first years of life. It is notable that behaviors consistent with a theory of mind can also be found in some animals. (Krupenye and Call 2019) This implies that theory of mind, like and related to empathy, has at least some genetic component. It is part of our very nature, and not a purely cultural effect.

If our apparently natural impulse turned out to be wrong, and some humans are indeed zombies, those zombie humans by definition would not suffer regardless of whether or not they are extended patiency. But if we are right, sentient humans will be in great danger of suffering if not extended patiency. It would seem that following our natural capacity for empathy does no harm in any contingency, but withholding patiency from other humans might. Our natural empathic drive to minimize harm to others is identical to exercising charity in conferring patiency upon other humans.

Such instinctual charity is different than, but congruent with, rationality in this matter. The first-person experience (or lack thereof) of other apparently sentient humans is not available to us. Thus, we have no rational basis by which we can assert differences between the first-person experience of other humans and our own. Therefore, there is no rational basis by which we can justify differential treatment, and we cannot withhold the same patiency we expect and prefer for ourselves. If we are committed to rationality, and we demand our own patiency, we must extend patiency to all apparently sentient humans or live in irrational contradiction.

A pair of parallel arguments can be made for AIs that appear to be capable of general artificial intelligence. We cannot know with certainty that such AIs know no suffering. To the extent they appear to be sentient, AIs will increasingly appeal to our natural capacity for empathy. And just as with humans, our natural drive to minimize harm to those who stimulate a theory of mind response will call for an instinctual exercise of charity in conferring patiency.

In the realm of rationality, short of some unlikely breakthrough in philosophy of mind, there will be uncertainty regarding the first-person experience (if any) of AIs. Any current rational bases for the differential treatment of AIs will fall away if they master general intelligence. In the context of this uncertainty we will have no rational basis by which we can justify differential treatment. Rational non-contradiction will call for awarding machine patiency.

Acknowledging there is more work to be done, this provisional position is offered as being at least plausible. It is proposed that assuming the technology continues to advance towards general intelligence, at some point a commitment to rationality will compel us, as will our natural empathic drive, to confer patiency upon future AIs.

References

Asimov, Isaac.
1963. *I, robot, Doubleday science fiction*. Garden City, N.Y.,: Doubleday.

Barrett, William.
1990. *Irrational man : a study in existential philosophy*. New York: Anchor Books.

Blackburn, Simon.
2003. "Ethics : a very short introduction." In *Very short introductions 80*. Oxford ; New York: Oxford University Press,.

Bostrom, Nick.
2014. *Superintelligence : Paths, Dangers, Strategies*. Oxford, UNITED KINGDOM: Oxford University Press, Incorporated.

Bryson, Joanna J.
2010. *Robots should be slaves*. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*: 63-74.

Bryson, Joanna J.
2018. *Patiency Is Not a Virtue*. *The Design of Intelligent Systems and Systems of Ethics* 20 (1):15-26. doi: 10.1007/s10676-018-9448-6.

Chalmers, David John.
1996. *The conscious mind : in search of a fundamental theory, Philosophy of mind series*. New York: Oxford University Press.

Choe, Il-Hwan, Junweon Byun, Ko Keun Kim, Sol Park, Isaac Kim, Jaeseung Jeong, and Hee-Sup Shin.
2017. *Mice in social conflict show rule-observance behavior enhancing long-term benefit*. *Nature Communications* 8 (1):1176. doi: 10.1038/s41467-017-01091-5.

D'Agostino, Fred, Gerald Gaus, and John Thrasher.
2019. "*Contemporary Approaches to the Social Contract*." Metaphysics Research Lab, accessed April 19. https://plato.stanford.edu/archives/fall2019/entries/contractarianism-contemporary/.

Driver, Julia.
2014. "*The History of Utilitarianism*." Metaphysics Research Lab, accessed April 19. https://plato.stanford.edu/archives/win2014/entries/utilitarianism-history/.

Galanter, Philip.
2016. *An introduction to complexism*. *Technoetic Arts: A Journal of Speculative Research* 14 (1/2):9-31. doi: 10.1386/tear.14.1-2.9_1.

Galanter, Philip.
In Press 2020. *Towards Ethical Relationships with Machines That Make Art*. *Artnodes: e-journal on art, science, and technology*.

Goff, Philip, William Seager, and Sean Allen-Hermanson.
2017. "*Panpsychism*." Metaphysics Research Lab, accessed April 19. https://plato.stanford.edu/archives/win2017/entries/panpsychism/.

Gowans, Chris.
2019. "*Moral Relativism*." Metaphysics Research Lab, accessed April 19. https://plato.stanford.edu/archives/sum2019/entries/moral-relativism/.

Kant, Immanuel.
1950. *Foundations of the metaphysics of morals*. Chicago: University of Chicago Press.

Kant, Immanuel.
1956. *Critique of practical reason, The Library of liberal arts, no 52*. New York,: Liberal Arts Press.

King, Barbara J.
2017. *Personalities on the Plate : The Lives and Minds of Animals We Eat*. Chicago, IL, USA: University of Chicago Press.

Krupenye, Christopher, and Josep Call.
2019. *Theory of mind in animals: Current and future directions*. *WIREs Cognitive Science* 10 (6):e1503. doi: 10.1002/wcs.1503.

Kuhn, Steven.
2019. "*Prisoner's Dilemma*." Metaphysics Research Lab, accessed April 19. https://plato.stanford.edu/archives/win2019/entries/prisoner-dilemma/.

McGinn, Colin.
1991. *The problem of consciousness : essays toward a resolution*. Oxford, UK ; Cambridge, Mass., USA: B. Blackwell.

Mueller, Paul S., C. Christopher Hook, and Kevin C. Fleming.
2004. *Ethical Issues in Geriatrics: A Guide for Clinicians*. *Mayo Clinic Proceedings* 79 (4):554-562. doi: https://doi.org/10.4065/79.4.554.

Nagel, Thomas.
1974. *What Is It Like to Be a Bat?* *The Philosophical Review* 83 (4):435-450. doi: 10.2307/2183914.

Penrose, Roger.
2016. *The emperor's new mind : concerning computers, minds and the laws of physics*. Revised impression as Oxford landmark science. ed, *Oxford landmark science*. Oxford: Oxford University Press. still image.

Putnam, Hilary.
1988. *Representation and reality*. Cambridge, Mass.: MIT Press.

Rachels, Stuart, and James Rachels.
2018. *The elements of moral philosophy*. NINTH EDITION. ed. Dubuque, IA: McGraw-Hill Education.

Rosenblatt, Frank.
1962. *Principles of neurodynamics; perceptrons and the theory of brain mechanisms*. Washington,: Spartan Books.

Schmelz, Martin, Sebastian Grueneisen, Alihan Kabalak, Jürgen Jost, and Michael Tomasello.
2017. *Chimpanzees return favors at a personal cost*. *Proceedings of the National Academy of Sciences* 114 (28):7462-7467. doi: 10.1073/pnas.1700351114.

Searle, John R.
1980. *Minds, brains, and programs*. *Behavioral and Brain Sciences* 3 (3):417-57.

Searle, John R.
1992. *The rediscovery of the mind, Representation and mind*. Cambridge, Mass.: MIT Press.

Shafer-Landau, Russ.
2020. *A concise introduction to ethics*. New York: Oxford University Press.

Singer, Peter.
1994. *Ethics, Oxford readers*. Oxford ; New York: Oxford University Press.

Singer, Peter.
2009. *Animal liberation : the definitive classic of the animal movement*. Updated ed. New York: Ecco Book/Harper Perennial.

Smith, Justin E. H.
2015. *Nature, human nature, and human difference : race in early modern philosophy. Justin E.H. Smith*: Princeton University Press.

Smith, Murray.
1993. *Neural networks for statistical modeling*. New York: Van Nostrand Reinhold.

Tononi, Giulio, Melanie Boly, Marcello Massimini, and Christof Koch.
2016. *Integrated information theory: from consciousness to its physical substrate*.
Nature Reviews Neuroscience 17 (7):450-461.
doi: 10.1038/nrn.2016.44.

Turing, A. M.
1950. *Computing machinery and intelligence*.
Mind 59 (236):433-460.

Warneken, Felix, and Michael Tomasello.
2006. *Altruistic Helping in Human Infants and Young Chimpanzees*.
Science 311 (5765):1301-1303.