The Lion, the Bat & the Thermostat: Metaphors on Consciousness

Brian L. Frye[1]

Abstract

Can robots have rights? It depends on the meaning of "robots" and "rights." Different kinds of robots can have different kinds of rights. Robots can already have the rights of things, and may soon be able to have the rights of legal entities. But it is unclear whether robots can have the rights of animals or persons. It probably depends on what theory of mind is true: dualist, reductionist, or agnostic. Under a dualist theory, robots can have rights if they possess a soul or other form of mental substance. Under a reductionist theory, robots can have rights if they are conscious, or at least functionally identical to a human or animal. And under an agnostic theory, it depends on how brains actually work.

Philosophers often use metaphors to explore problems they do not understand. As Thomas Nagel wryly observed, "philosophers share the general human weakness for explanations of what is incomprehensible in terms suited for what is familiar and well understood, though entirely different."[2] Philosophers of mind are no exception, and often rely on three popular metaphors: the lion, the bat, and the thermostat. I will describe these metaphors, and reflect on how they may illuminate our speculations on the possibility of artificial intelligence.

Introduction

"But we cannot reckon with what is lost when we start out to transform the world. Man shall be free and supreme; he shall have no other aim, no other labor, no other care than to perfect himself. He shall serve neither matter nor man. He will not be a machine and a device for production. He will be Lord of creation."[3]

On October 25, 2017, Saudi Arabia granted "citizenship" to a humanoid robot named "Sophia."[4] Obviously, Sophia's "naturalization" was purely symbolic, as "she" is an object that manifests only an amusing simulacrum of personhood. But many still found it ironic, as Saudi Arabia discriminates against women and denies citizenship to most foreign residents. Nevertheless,

---

[1] Spears-Gilbert Associate Professor of Law, University of Kentucky School of Law. J.D., New York University School of Law, 2005; M.F.A., San Francisco Art Institute, 1997; B.A, University of California, Berkeley, 1995. Thanks to Patrick S. O'Donnell and Katrina Dixon for their helpful suggestions.

[2] Thomas Nagel, What is it like to be a bat?, 83 Phil. Rev. 4 (1974).

[3] Karel Čapek, R.U.R. (Rossum's Universal Robots) (Trans. Paul Selver & Nigel Playfair 1920), available at http://preprints.readingroo.ms/RUR/rur.pdf.

[4] Cleve R. Wootson Jr., Saudi Arabia, which denies women equal rights, makes a robot a citizen, Washington Post, October 29, 2017; Cristina Maza, Saudi Arabia Gives Citizenship to a Non-Muslim, English-Speaking Robot, Newsweek, October 26, 2017; Tracy Alloway, Saudi Arabia Gives Citizenship to a Robot, Bloomberg, October 26, 2017.

even Sophia's ersatz citizenship encourages speculation as to whether a robot could actually become a citizen and exercise the rights of citizenship.

I. Introduction

"I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'."[5]

In his 1992 article, *Legal Personhood for Artificial Intelligences*, Lawrence Solum explored the possibility of "robot rights" by asking both whether an artificial intelligence could become a legal person. More specifically, he asked whether an artificial intelligence could serve as a trustee, and whether an artificial intelligence could claim the constitutional rights of a natural person.[6] He asked these pragmatic legal questions in order to investigate the practical consequences of philosophical claims relating to the possibility of artificial intelligence.

Solum concluded that an artificial intelligence capable of passing the Turing Test probably could serve as a trustee, because it could insure against negligence, could not engage in conduct requiring deterrence, and could exercise judgment sufficient to administer most trusts under most circumstances. Likewise, he argued that an artificial intelligence capable of passing the Turing Test could and should be entitled to claim at least some of the constitutional rights of a natural person.

While Solum acknowledged the various philosophical theories of mind, he concluded that they were largely irrelevant to the legal questions at issue. He observed that if an artificial intelligence could simulate a human being in all relevant respects, it probably would not matter if the artificial intelligence lacked "consciousness," however defined. "My prediction (and it is only that) is that the lack of real intentionality would not make much difference if it became useful for us to treat AIs as intentional systems in our daily lives."[7]

Solum was probably right. If an artificial intelligence were indistinguishable from a human being in all relevant respects, courts probably would and should allow it to assert the legal rights of a human being. Of course, if science develops an empirical theory of consciousness, then courts probably would not and should not allow an artificial intelligence that merely simulates consciousness, without actually being conscious, to assert those rights. But in the absence of such an empirical theory, courts would probably dismiss concerns about the possibility of a "philosophical zombie."

But what if Solum's questions were considered from the opposite perspective? He asked whether courts would and should allow an artificial intelligence that replicates a human being to assert legal rights, and avoided philosophical questions about the nature of consciousness. By

---

[5] Alam M. Turing, Computing Machinery and Intelligence, 59 Mind 433 (1950).
[6] Lawrence B. Solum, Legal Personhood for Artificial Intelligences, 70 N.C. L. Rev. 1231 (1992).
[7] Lawrence B. Solum, Legal Personhood for Artificial Intelligences, 70 N.C. L. Rev. 1231, 1282 (1992).

contrast, I will ask what different theories of mind imply about whether and when an artificial intelligence could and should be allowed to assert the legal rights of a natural person.

Can robots have rights? It depends on the meaning of "robots" and "rights." Different kinds of robots can have different kinds of rights. Robots can already have the rights of things, and may soon be able to have the rights of legal entities. But it is unclear whether robots can have the rights of animals or persons. It probably depends on what theory of mind is true: dualist, reductionist, or agnostic. Under a dualist theory, robots can have rights if they possess a soul or other form of mental substance. Under a reductionist theory, robots can have rights if they are conscious, or at least functionally identical to a human or animal. And under an mysterian theory, it depends on how brains actually work.

Setting aside formal theories of mind, how do we actually think about consciousness? Philosophers often use metaphors to explore problems they do not understand. As Thomas Nagel wryly observed, "philosophers share the general human weakness for explanations of what is incomprehensible in terms suited for what is familiar and well understood, though entirely different."[8] Philosophers of mind are no exception, and often rely on three popular metaphors: the lion, the bat, and the thermostat. I will describe these metaphors, and reflect on how they may illuminate our speculations on the possibility of artificial intelligence.

II. What is a Robot?

"That man is the noblest creature may also be inferred from the fact that no other creature has yet contested this claim."[9]

The word "robot" was coined in 1920 by the Czech writer and painter Josef Čapek, for use in his brother Karel Čapek's play *R.U.R. (Rossum's Universal Robots)*.[10] The Czech word "robota" means "serf" or "compulsory labor." In Čapek's play, a factory manufactures artificial people or "robots" for use as slaves. Initially, the robots lack consciousness and are docile, but some of the robots become self-aware and incite a robot rebellion, which eventually leads to the extinction of humanity.

Today, a "robot" is typically defined as "a machine capable of carrying out a complex series of actions automatically."[11] Robots take many forms and can perform a wide variety of tasks. Most robots consist of a machine controlled by a computer, either directly or remotely. As robots become increasingly sophisticated, they can supplement or replace human labor in many

---

[8] Thomas Nagel, What is it like to be a bat?, 83 Phil. Rev. 4 (1974).
[9] Georg Christoph Lichtenberg, The Waste-Books, D 58 (1776).
[10] Karel Čapek, About the Word Robot, Lidove noviny, December 24, 1933, available at https://web.archive.org/web/20130123023343/http://capek.misto.cz/english/robot.html ("'But,' the author said, 'I don't know what to call these artificial workers. I could call them Labori, but that strikes me as a bit bookish.' 'Then call them Robots,' the painter muttered, brush in mouth, and went on painting.").
[11] Robot, Oxford English Dictionary.

contexts. Industrial robots substitute for human labor in factories, domestic robots substitute for human labor in and around homes, and military robots substitute for human labor on the battlefield. Most robots are designed to perform a particular task or set of tasks. But some humanoid robots are designed to resemble a human and replicate human behavior, to a greater or lesser degree. Currently, robots can resemble humans closely enough to be creepy (the "uncanny valley"), but not closely enough to fool anyone, and robots can only replicate formulaic human behaviors. They certainly cannot replicate human thought.

Often unwittingly, we use robots and artificial intelligence as metaphors to discuss uncomfortable ideas. The thinking robot is a "hypothetical" that enables a form of cathexis. Ironically, while Capek explicitly used his robots as metaphors for every social evil under the sun - slavery, class warfare, nationalism, and so on - today many of us seem to forget that our "thinking robots" are metaphors at all.

For example, science-fiction authors often imagine robots that replicate humans, in appearance or behavior. These imaginary robots typically possess "artificial intelligence," meaning that they are conscious, or at least exhibit the objective indicia of consciousness, like the ability to engage in abstract reasoning and communicate with humans in natural language. In other words, they act as if they were humans, albeit often curiously affectless or otherwise peculiar ones.

A cynic might observe that science-fiction's stock artificially intelligent robot is resembles Andy Kaufman's character Latka from the sitcom *Taxi*: a person who fails to behave or respond to social cues in the expected way. This should come as no surprise, as science-fiction has always used robots as metaphors for the Other, a comfortable way to ask uncomfortable questions, typically whether and how difference can be assimilated.

A. What is "Artificial Intelligence"?

"Artificial intelligence" is increasingly a buzzword in computer science and other fields. In the broadest sense, artificial intelligence means a human-made machine that can replicate a cognitive function of a human. The idea of artificial intelligence has existed since time immemorial, from Pygmalion's living statue of Galatea and Hephaestus's bronze man Talos, to the Golem of Jewish legend and Yan Shi's automaton, to King Ajatashatru of Magadha's mechanical guards and Rocail's autonomous statutes. But it became salient in the 20th century, when we invented computers. And it left the realm of science-fiction in the 21st century, when computers began to compete with and replace humans in the knowledge economy.

But the meaning of the term "artificial intelligence" still depends on the context in which it is used. Specifically, scholars and researchers typically distinguish between "weak" and "strong" artificial intelligence. The terms are not intended as an assessment of the value of different forms of artificial intelligence, but rather a description of how they work and what they are intended to accomplish. Indeed, "weak" artificial intelligence is rapidly transforming contemporary society, while "strong" artificial intelligence remains a pipe dream.

B. "Weak" Artificial Intelligence

"Weak" artificial intelligence is the ability to replicate a cognitive function of a human by any means. Specifically, weak artificial intelligence does not require consciousness. Ironically, weak artificial intelligence has proven incredibly powerful. In particular, machine learning and other techniques have enabled computers to accomplish many tasks that previously required human intervention, often more efficiently than humans can accomplish those tasks. While weak artificial intelligence presents certain normative concerns, especially when it relies on "black box" algorithms that cannot be reverse-engineered, it has already begun to transform contemporary society, and surely will only become more important.

C. "Strong" Artificial Intelligence

"Strong" artificial intelligence is a human-made conscious entity. Of course, it is trivially true that strong artificial intelligence is possible. Every human is a strong artificial intelligence, and we create a strong artificial intelligence every time we reproduce. Indeed, almost every living thing is a strong artificial intelligence, or at least a step on the way to strong artificial intelligence. The world tells us not only that strong artificial intelligence is possible, but also that it can take many forms. In other words, we know how to create strong artificial intelligence, but we don't know how we create it. Or rather, we know that strong artificial intelligence is possible, but we don't know why, or how it works.

While computer science has many successes in creating weak artificial intelligence, it has no successes in creating strong artificial intelligence. We have created computers that can solve all manner of different problems, and surely will create computers that can solve ever more problems, but we cannot create computers that are conscious, or even smack of consciousness. Our computers can do many things, but they cannot reproduce the behavior of even the simplest biological organism, let alone a human being.

For example, scientists have learned an immense amount about and from the roundworm *C. elegans*.[12] They have even mapped all 302 neurons in its brain. But they still do not know how its brain actually works. Indeed, they cannot even model the functionality of this simplest of brains. Surely, neurologists will eventually solve the puzzle of the roundworm brain, and with it every other brain. But they are not yet close. Indeed, it is not yet clear that they even know what questions to ask.

III. What Are Rights?

[12] See, e.g., Sydney Brenner, Nature's Gift to Science, Nobel Lecture, December 8, 2002, available at https://www.nobelprize.org/nobel_prizes/medicine/laureates/2002/brenner-lecture.pdf

Broadly defined, “Rights are entitlements (not) to perform certain actions, or (not) to be in certain states; or entitlements that others (not) perform certain actions or (not) be in certain states.”[13] For the purpose of this article, I will consider only “legal rights,” or “rights which exist under the rules of legal systems or by virtue of decisions of suitably authoritative bodies within them.”[14] Different definitions of rights provide that different kinds of entities can have legal rights. If rights protect interests, then any entity that would benefit from another’s performance of a legal duty can have legal rights, but if rights protect choices, then only entities that can perform a legal duty can have legal rights.

In considering the possibility of “robot rights,” it may be helpful to identify different categories of entities, and the kinds of rights that each of those categories of entities may hold. The kinds of rights a robot can have will depend on which categories of entity a robot can occupy: non-person or person, thing or animal, legal person or natural person.

A. Rights of Non-persons

Inanimate things and non-human animals can have interests, and therefore can possess certain kinds of rights in certain circumstances.[15] Historically, temples and churches could possess certain legal rights. The law of deodands provided that chattel property and animals could be held legally responsible for a person’s death.[16] In rem jurisdiction provides that chattel property can be the subject of legal proceeding. And the law often provides certain rights and protections to non-human animals.[17] Of course, non-persons cannot themselves assert legal rights, and must rely on others to assert those rights on their behalf.

1. Rights of Things

The law protects certain inanimate things that may benefit from legal protection. Indeed, one could view property law as the law of the rights of things and how those rights may be asserted. From a law and economics perspective, we observe that protecting the “rights” of rivalrous things may increase their contribution to net social welfare, by reducing transaction costs and facilitating efficient uses.[18] And from a normative perspective, we observe that the law often protects the “rights” of certain inanimate things in order to recognize and express their social value. Historic preservation laws protect certain historic and archaeological sites from certain

---

[13] Rights, Stanford Encyclopedia of Philosophy, at https://plato.stanford.edu/entries/rights/
[14] Legal Rights, Stanford Encyclopedia of Philosophy, at https://plato.stanford.edu/entries/legal-rights/
[15] See generally, John Chipman Gray, The Nature and Sources of the Law (1909)(1921).
[16] See, e.g., Anna Pervukhin, Deodands: A Study in the Creation of Common Law Rules, 47 Am. J. L. Hist. 237 (2005). In theory, the law of deodands could be revived and applied to robots. See, e.g., Christina Mulligan, Revenge Against Robots, South Carolina Law Review, Forthcoming, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3016048
[17] See, e.g., Joyce Tischler, The History of Animal Law, Part I (1972–1987), 1 Stan. J. Animal L. & Pol’y 3–12 (2008), available at http://sjalp.stanford.edu/pdfs/Tischler.pdf and Anthony D'Amato & Sudhir K. Chopra, Whales: Their Emerging Right to Life, 85 Am. J. Int’l L. 21 (1991).
[18] Cf. Jeremy Kidd, Kindergarten Coase, 17 Green Bag 2d. 141 (2014).

kinds of uses that will reduce their social value.[19] Moral rights laws protect certain works of authorship from alteration or destruction.[20] And environmental laws protect certain ecosystems from alteration or destruction.[21] Legal scholars have argued that those rights should extend more broadly to ensure and facilitate the protection of socially valuable things.[22] Surely, some or all of the forms of legal rights granted to things could be extended to robots as well, if we believed that it would be socially beneficial to do so.

2. Rights of Animals

The law also protects certain animals under certain circumstances.[23] For example, many laws prohibit the abuse of domestic animals. Other laws prohibit certain forms of abuse of livestock. And still other laws protect certain forms of wildlife in a variety of ways and for a variety of reasons.[24] Some scholars have argued that similar rights should extend to all living things.[25] Of course, animals and other living things cannot assert rights on their own behalf, any more than things. But it is unclear whether robots could have the kinds of legal rights given to animals and other living things, because it is unclear whether robots could ever become the moral equivalent of an animal or living thing. In any case, they certainly are not today, although people may impute moral significance to robot "welfare."

B. Rights of Persons

Only persons can assert legal rights on their own behalf. But personhood has different meanings in different contexts, and grants different rights to different categories of persons.[26] How should we determine what kinds of entities can qualify as persons and what rights each category of persons can claim?

The word "person" has both a colloquial and legal meaning. A colloquial "person" is any human being, but a legal "person" is any subject of legal rights.[27] Accordingly a legal person can be a

---

[19] See, e.g., National Historic Preservation Act of 1966, 80 Stat. 915.
[20] See, e.g., The Visual Artists Rights Act of 1990 (VARA), 17 U.S.C. § 106A and the Berne Convention, Article 6bis.
[21] See, e.g., North American Wetlands Conservation Act, (P.L. 101-233) (December 13, 1989).
[22] See, e.g., Christopher Stone, Earth and Other Ethics (1987).
[23] See, e.g., Cass Sunstein & Martha Nussbaum, Animal Rights: Current Debates and New Directions (2005).
[24] See, e.g., The Migratory Bird Treaty Act of 1918, 16 U.S.C. §§ 703 et seq.
[25] See, e.g., Christopher Stone, Should Trees Have Standing?: Toward Legal Rights for Natural Objects, 45 S. Cal. L. Rev. 450 (1972).
[26] See generally Zoe Robinson, Constitutional Personhood, 84 George Washington Law Review 605 (2016). Note, What We Talk About When We Talk About Persons: The Language of a Legal Fiction, 114 Harv. L. Rev. 1745 (2001).
[27] Lawrence B. Solum, Legal Theory Lexicon 027: Persons and Personhood, LEGAL THEORY LEXICON (Mar. 14, 2004) http://lsolum.typepad.com/legal_theory_lexicon/2004/03/legal_theory_le_2.html. See also Note, What We Talk About When We Talk About Persons: The Language of a Legal

“natural person,” i.e., a human, or a “juridical person,” i.e., a legal entity that can have certain legal rights. Courts have long held that corporations - and more recently limited liability companies - are legal persons, entitled to certain legal rights. For example, the Supreme Court recently held that corporate persons can assert the right to free speech under the First Amendment.[28]

How should we determine whether an artificial intelligence can have particular rights? A functionalist approach would determine the purpose of the right and whether extending it would further that purpose.[29] In other words, the existence of a right would depend on both the nature of the right and the nature of the claimant.[30]

It seems like a different question to ask whether an artificial intelligence could have the rights of a legal entity versus the rights of a natural person. It is true that one reason legal entities can have some of the rights of a constitutional person is because they are ultimately controlled by natural persons. A legal entity does not and cannot make decisions on its own. Its decisions are made by one or more natural persons acting as its agents. But many entities do not have some of the rights of constitutional persons, like partnerships and trusts. This is a purely functional question. Sometimes it makes sense to extend constitutional personhood to a particular kind of legal entity under circumstances, and sometimes it does not.

But it could never be the case that a legal entity could become a natural person. Legal entities are a means to an end. Natural persons are ends in themselves. The purpose of a legal entity to enable natural persons to achieve their ends more efficiently.

Ultimately, whether an artificial intelligence can be a natural person is a fundamentally different question. Currently, we assume that robots are things. Is it possible that an artificially intelligent robot could be a natural person? Or will robots always be something other than a natural person? And if so, how should we categorize them? It seems that the answer must depend on whether artificial intelligence is possible, and if so, what it will look like.

To put it differently, it seems unproblematic that a robot could have legal personhood, in the way that a legal entity can have legal personhood. Accordingly, as Solum recognizes, a robot’s actions could be legally binding. This does not seem fundamentally different than holding that the actions of a legal entity can be binding. The only question is when? While that is a difficult

---

Fiction, 114 Harv. L. Rev. 1745 (2001) (quoting JOHN CHIPMAN GRAY, THE NATURE AND SOURCES OF THE LAW 27 (Roland Gary ed., 2d ed. 1931) (1909)) (“John Chipman Gray observed that "[i]n books of Law, as in other books, and in common speech, 'person' is often used as meaning a human being, but the technical legal meaning of a 'person' is a subject of legal rights and duties."”).

[28] Citizens United v. Federal Election Commission 558 U.S. 310 (2010).

[29] Zoe Robinson, Constitutional Personhood, 84 George Washington Law Review 605, 661 (2016).

[30] Zoe Robinson, Constitutional Personhood, 84 George Washington Law Review 605, 661 (2016).

question, it is no more difficult than the question of when a corporation should be treated as a legal person.[31] We do not have a coherent theory of corporate personhood either.

Likewise, robots could be protected as a separate category, distinct from legal persons. For example, slaves were often protected as a separate legal category, in order to avoid the question of whether a slave was a "person" under the law.[32] And today, fetuses are protected as a separate category, because the Supreme Court has held that a fetus is not a person.[33]

Questions about whether rights should attach necessarily implicate conclusions about the purpose of those rights. In the case of robots, the question is how they are properly characterized. Or rather the question "can robots have rights" is a metaphysical question about what theory of mind is actually true. But "should robots have rights" is a political question about who will be included in the political community.

Why does it matter? Because the "person" metaphor contains considerable expressive content. When courts denied the personhood of slaves, they often expressed considerable reluctance.[34] And when the Supreme Court denied the personhood of a fetus, it did so only with hesitation and uncertainty.[35] Ultimately, the question of "robot rights" is fraught only because we use the "robot" metaphor to ask uncomfortable questions about the "person" metaphor. Asking questions about "robot rights" allows us to ask sublimated questions about who qualifies for legal personhood and why.

1. Rights of Juridical Persons

The law enables certain legal entities to assert certain legal rights as legal persons, specifically corporations and limited liability companies. Legal entities can have a wide range of legal rights, but cannot have all of the legal rights associated with natural persons. For example, a corporation is a legal person for the purpose of the First Amendment right of free speech, the Fourth Amendment rights to freedom from search and seizure and privacy, the Fifth Amendment right to freedom from double jeopardy and takings, the Sixth amendment right to counsel and trial by jury, and the Fourteenth Amendment rights to equal protection and due process. But a

---

[31] Zoe Robinson, Constitutional Personhood, 84 George Washington Law Review 605, 663 (2016) ("Answering the question of whether a corporation fulfills the identified function of a right requires a clear conception of corporate personhood.").
[32] Note, What We Talk About When We Talk About Persons: The Language of a Legal Fiction, 114 Harv. L. Rev. 1745, 1747-50 (2001).
[33] Note, What We Talk About When We Talk About Persons: The Language of a Legal Fiction, 114 Harv. L. Rev. 1745, 1754-59 (2001).
[34] Note, What We Talk About When We Talk About Persons: The Language of a Legal Fiction, 114 Harv. L. Rev. 1745, 1760-62 (2001).
[35] Note, What We Talk About When We Talk About Persons: The Language of a Legal Fiction, 114 Harv. L. Rev. 1745, 1762-64 (2001).

corporation is not a legal person for the purpose of the Fifth Amendment right to freedom from self-incrimination, the right to appointed counsel, or the Privileges and Immunities Clause.[36]

The Supreme Court has not explicitly articulated a theory of when legal entities are constitutional persons, but it tends to take a functionalist approach, asking whether extending a particular right or duty to a corporation would further the purpose of that right or duty. For example, the Supreme Court found that the purpose of the First Amendment right to free speech is to protect the dissemination of speech from government interference, regardless of its source, so extending the right to legal entities was consistent with the purpose of the right.[37]

On occasion, the Supreme Court has engaged in a surprising amount of anthropomorphism in explaining its conclusion that certain legal rights extend to corporations.[38] For example, it found that the Fifth Amendment protection against double jeopardy extended to corporations in order to prevent "embarrassment," "anxiety," and "insecurity."[39] And it found that the Fourth Amendment protection from unreasonable searches extends to corporations based on their "reasonable . . . expectation of privacy."[40]

This is probably unhelpful. Whether legal entities can have certain legal rights should depend only on whether it is socially beneficial to allow them to assert those rights as legal entities. Imputing beliefs and feelings to a legal entity is nonsensical and unhelpful. The legal rights of a legal entity should depend on its function and purpose, and whether extending particular rights will further that function and purpose.

Likewise, whether robots can have the rights of a juridical person is a functional question. If a robot can accomplish the functions of a legal entity, then surely it can and probably should have the rights of a legal entity. After all, in a sense, a corporation is like an imaginary legal robot, responding to the various needs and desires of its shareholders and stakeholders, for whose benefit it exists.

2. Rights of Natural Persons

Natural persons or "humans" have the most legal rights. But not all natural persons have same legal rights, or can assert those rights on their own behalf. Some people have limited rights, like noncitizens and felons. And others have rights that can only be asserted by others, like children and mentally disabled persons.

---

[36]

[37] Citizens United.

[38] Note, What We Talk About When We Talk About Persons: The Language of a Legal Fiction, 114 Harv. L. Rev. 1745 (2001).

[39] United States v. Martin Linen Supply Co., 430 U.S. 564 (1977).

[40] Dow Chemical Co. v. United States, 476 U.S. 227 (1986).

While the Supreme Court typically assumes that natural persons are legal persons, it asks whether a particular rights extend to particular classes of natural persons under particular circumstances.[41] Accordingly, noncitizens are legal persons who lack certain rights, based on their citizenship status. And felons are legal persons who lack certain rights, based on their prior actions. In neither case is there a question about their ontological status, only their ability to assert certain legal claims.

While mentally disabled people cannot assert their own rights, they still have rights, which a surrogate can assert on their behalf.[42] Likewise, minors have rights, which may be asserted by their parents or some other surrogate.

Whether robots can have the rights of a natural person necessarily depends on which theory of mind is true. If a computer can be conscious, then a robot can have the rights of a natural person, and vice versa.

IV. The Philosophy of Mind

Philosophers have pondered the "mind-body" problem since time immemorial, and developed a vast congeries of metaphysical theories of mind. While the number and variety of those theories defies comprehensive summary, they broadly fall into two categories: dualism and physicalism. Essentially, dualist theories hold that mental states and physical states are ontologically distinct, and physicalist theories hold that mental states are just physical states. In other words, dualist theories hold that mind and body are separate, and physicalist theories hold that they are the same.

A. Dualism

Dualist theories of mind hold that the mind and the body are distinct and that mental states cannot be reduced to physical states.[43] Classical dualism distinguished between material bodies and immaterial forms, including the soul. Platonic dualism held that the soul is an immaterial form imprisoned in a material body, and Aristotelian dualism held that the soul is an immaterial form realized in a material body.[44]

Modern dualism began with Cartesian "substance" dualism, which held that there are two kinds of substance: matter, which has the essential property of physical extension; and mind, which has the essential property of thinking.[45] By contrast, "property" dualism holds that mind or

---

[41] Zoe Robinson, Constitutional Personhood, 84 George Washington Law Review 605, 633-34 (2016).
[42] Cruzan v. Director, Missouri Department of Health, 497 U.S. 261 (1990).
[43] Howard Robinson, Dualism, Stanford Encyclopedia of Philosophy (2016), at https://plato.stanford.edu/entries/dualism/
[44] See Plato, Phaedo, 78b4–84b8 and Aristotle, De Anima III,4; 429a10–b9.
[45] Descartes, Meditations.

consciousness is a non-physical property of physical substance, either a fundamental property of reality or an emergent property of complex physical systems.

Dualist theories of mind often rely on intuitions about subjective experiences or "qualia" and the conceivability of a mind existing without a body, and *vice versa*. For example, the "zombie hypothesis" relies on the conceivability of a "philosophical zombie" or person lacking consciousness as evidence of the mind-body distinction. The persistent objection to dualist theories of mind is that they cannot explain the existence of non-physical things, or how they interact with physical things.

Under a dualist theory of mind, it seems unproblematic that a robot could have the rights of a natural person if and only if it can possess an immaterial soul or mind that enables it to have mental states. If a robot cannot have mental states, it cannot be a subject of legal rights. But a mindless robot could have the rights of a thing or a legal entity.

B. Physicalism

Physicalist theories of mind hold that everything supervenes on the physical world, including the mind. While there are many different physicalist theories of mind, they broadly fall into two categories: reductionist and non-reductionist. Reductionist theories typically holds that mental states are just brain states. The dominant reductionist theory is the computational theory of mind, which holds that the brain is just a universal Turing machine, and that consciousness, however defined, can be realized on any universal Turing machine, including a computer. Non-reductionist theories are characterized more by what they question than what they claim. In general, they observe that we do not know how the brain works or what it means to say that a mental state is identical to a brain state, and therefore cannot make claims about the nature of consciousness or how it can be realized. I will refer to these as "agnostic" theories of mind, as they generally do not deny the computation theory of mind and its presumption of "multiple realizability," but assert the "Scotch verdict" of "not proven" and suggest that waiting for more evidence.

1. Turing Machines & the Turing Test

The English computer scientist, mathematician, logician, cryptanalyst, philosopher and theoretical biologist Alan Mathison Turing was one of the greatest thinkers of the 20th century. Among many other things, he conceived of the general purpose computer and invented computer science. In 1936, Turing explained how to create a machine that could be used to compute any computable sequence.[46] The machine he described became known as a "universal Turing machine," and provided the theoretical framework for the general purpose computer. A "Turing machine" is a mathematical model of computation that defines an abstract

[46] Alan M. Turing, On Computable Numbers, with an Application to the Entscheidungsproblem, 2 Proceedings of the London Mathematical Society 230 (1936).

machine, which manipulates symbols on a strip of tape according to a table of rules, and a universal Turing machine is a machine that can simulate any other Turing machine. John von Neumann built on Turing's ideas to propose the von Neumann architecture that is the basis of the modern computer.[47]

But Turing was also the father of of artificial intelligence. In 1950, he famously proposed a test to determine whether a computer could think.[48] Under the "Turing test" for artificial intelligence, a human must pose a series of questions, which are answered by both a human and a computer. If the human questioner cannot distinguish between the human and computer answers, then the computer has passed the Turing test.

Essentially, Turing reframed the question "Can machines think?" as "Can machines imitate thinking?" And in so doing, he implicitly proposed a computational theory of mind. Interestingly, the Turing test for artificial intelligence was anticipated in the negative by Rene Descartes in the 17th century:

> If there were machines which bore a resemblance to our bodies and imitated our actions as closely as possible for all practical purposes, we should still have two very certain means of recognizing that they were not real men. The first is that they could never use words, or put together signs, as we do in order to declare our thoughts to others. For we can certainly conceive of a machine so constructed that it utters words, and even utters words that correspond to bodily actions causing a change in its organs. … But it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of men can do.[49]

While Turing disclaimed any intention to question the "mystery" of consciousness, he suggested that "thinking" may not require consciousness. "I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper."[50]

Of course, there are many objections to the Turing test for artificial intelligence, many of which were anticipated and addressed in the paper proposing the test.[51] Critics have argued that it is too easy and too hard; too narrow and too broad; neither logically necessary nor sufficient;

---

[47] For a comprehensive and comprehensible discussion of Turing and von Neumann's contributions to computer science and artificial intelligence, see Jeffrey M. Lipshaw, Halting, Intuition, Heuristics, and Action: Alan Turing and the Theoretical Constraints on AI-Lawyering, Savannah Law Review, Forthcoming

[48] Alan M. Turing, Computing Machinery and Intelligence, 59 Mind 433 (1950).

[49] Rene Descartes, The Discourse on the Method (1637).

[50] Alan M. Turing, Computing Machinery and Intelligence, 59 Mind 433 (1950).

[51] Alan M. Turing, Computing Machinery and Intelligence, 59 Mind 433 (1950).

dangerous or harmful; and defeasible or merely probabilistic.[52] The strength of the Turing test is that it proposes a functional account of intelligence, under which intelligence is simply whatever produces the appearance of intelligence.

2. Reductionism

Today, the prevailing theory of mind is the computational theory, which holds that a mind is simply a "universal Turing machine" or "computational system": an information processing machine that can perform any calculation.[53] Hilary Putnam introduced the computational theory of mind in 1967, and it soon became the dominant theory of mind.[54] The classical computational theory holds that mental states are functional states that are multiply realizable, and that a system has a mind if it has a suitable functional organization. There are several variations on the computational theory. For example, the representational theory holds that mental states are computational, but not necessarily functional. And connectionist theories holds that mental states are computational, but in the form of a neural network, not a Turing machine.

Proponents of the computational theory of mind take different positions on intentionality, or the "aboutness" of mental states, ranging from intentional realism, which holds that intentionality is a real property of mental states, to eliminativism, which holds that it is not. Some philosophers stake out a middle ground, like Daniel Dennett and Donald Davidson, who observe that the concept of intentionality is useful, but imply that it may not be a real property of mental states.[55]

There are many criticisms of the computational theory of mind, including triviality, the "incompleteness" of Turing machines, functional limitations on computational modeling, its underspecification of temporality, and its failure to account for interactions between the mind and body. But the best-known objection to the computational theory of mind is the "Chinese Room" argument.

Under reductionist theories of mind, including the computational theory of mind, it seems unproblematic that a robot could have the rights of a natural person. After all, under the computational theory of mind, a brain is just a computational system, so an artificially intelligent robot is functionally identical to a natural person. Of course, an non-artificially intelligent robot could have the rights of a thing or a legal entity.

3. The Chinese Room Argument

---

[52] The Turing Test, Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/entries/turing-test/
[53] Michael Rescorla, The Computational Theory of Mind, Stanford Encyclopedia of Philosophy, at https://plato.stanford.edu/entries/computational-mind/
[54] Hilary Putnam, Psychophysical Predicates, in Art, Mind, and Religion (W. Capitan & D. Merrill eds. (1967)).
[55] See generally, Daniel Dennett, The Intentional Stance (1987) and Donald Davidson, Essays on Actions and Events (1980).

In 1981, John Searle responded to the computational theory of mind with the “Chinese Room” argument, a thought experiment intended to show that it is impossible for a digital computer to be conscious.[56] Essentially, Searle imagines an English-speaking person in a sealed room, which contains instructions in English for how to manipulate Chinese characters. The person receives Chinese phrases, follows the instructions, and provides Chinese phrases in response. To someone outside the room, the Chinese Room appears to understand Chinese. But the person in the room does not actually understand Chinese.

The Chinese Room is analogous to a computer. Just like the person in the Chinese Room simply manipulates symbols based on predetermined rules, a computer simply manipulates data based on predetermined rules. The person in the Chinese Room does not understand Chinese, and does not need to understand Chinese. Likewise, a computer does not understand the meaning of the data it manipulates, and does not need to understand it. Both follow instructions that create the appearance of understanding the significance of the data they receive and produce, without actually understanding it. To put it another way, they manipulate syntax without understanding semantics.

Essentially, the Chinese Room argument is an inversion of the Turing test. Turing argued that the best test for consciousness is whether a computer acts like it is conscious. Searle responded that consciousness is a property of the mind that either exists or does not.

The point of the Chinese Room argument is to refute functional theories of mind. Searle argues that the point of a theory of mind is to explain consciousness, but functional theories of mind just explain it away, essentially by saying it doesn’t exist or doesn’t matter. Specifically, computational theories hold that the mind is just a computational system. But Searle observed that they provide no theory of how a computer could become conscious, or even try to account for consciousness or subjective experience.

In addition, Searle observed that proponents of the computational theory of mind are arguably crypto-dualists. Their assertion that the mind is multiply realizable, a kind of “software” than can run on the “hardware” of the brain or any other adequate machine, implies that the mind exists independently of the physical brain.

One common response to the Chinese Room argument is that while the person in the room doesn’t understand Chinese, the system as a whole understands Chinese. Accordingly, while the CPU of a computer may not understand the meaning of the data it manipulates, a computing system as a whole could understand it. Specifically, perhaps semantics is simply an emergent property of complex manipulation of syntax, or rather, perhaps meaning is an emergent property of complex computation.

---

[56] John R. Searle, Minds, Brains, and Programs, 3 Behavioral and Brain Sciences 417 (1981). See also John R. Searle, Minds, Brains and Science 28 (1984). See also The Chinese Room Argument, Stanford Encyclopedia of Philosophy, at https://plato.stanford.edu/entries/chinese-room/#3

Another common response to imagine a “brain simulator,” or machine that duplicates on a computer every physical process that occurs in a brain. Presumably such a simulator would be “conscious” in the same way as a brain. Searle has responded that a simulation of consciousness is not the real thing. If a computer simulates a hurricane, it can predict what the hurricane will do, but it cannot create a hurricane. Likewise, if a computer can simulate a brain, presumably it could predict what the brain will do, but it cannot create a brain. We don’t know what causes consciousness, but we do know that following instructions doesn’t cause consciousness. Or rather, the appearance of intentionality is not the same thing as the presence of intentionality.

More fundamentally, critics argue that the Chinese Room argument relies on intuitions about understanding and consciousness that are unreliable and inconsistent with scientific knowledge. For example, Daniel Dennett has dismissed the Chinese Room argument as an “intuition pump,” which draws fallacious conclusions from a misleading thought experiment. Many supporters of the computational theory of mind hold that the missing factor is speed, and that when computers become fast enough, we will see that brains are just fast computational systems. Searle’s implicit response is that the computational theory of mind also relies on intuitions about how brains work, and that it is at least equally likely that the reason the computational theory of mind is false is that brains do things we don’t yet understand, and digital computers can’t do those thing.

4. Mysterianism

While Searle and other philosophers argue that the computational theory of mind is wrong, they still believe that physicalism is true. Typically, they argue that the physical properties of the brain cause consciousness, but we don’t know how, and don’t have any reason to believe that the physical properties of a computer can cause consciousness. In other words, living things are “biological machines” and consciousness is a “biological phenomenon.” If and when we come to understand how biological processes can cause consciousness, then we may be able to determine whether non-biological processes can also cause consciousness. But we currently do not have the capacity to answer that question. It is a mystery. Some “mysterians” believe that humans will never learn how brains works, because cannot understand consciousness. Others believe that humans cannot understand how brains work now, but may learn how to answer the question in the future.

Under a mysterian theory of mind, whether a robot could have the rights of a natural person depends on it turns out that consciousness actually works. Presumably, a biological robot could have the rights of a natural person, if it were conscious. But a non-biological robot could have the rights of a natural person only if it turns out that non-biological machines can become conscious. As it stands, the question is unanswerable, because we do not know what causes consciousness or how brains cause consciousness. Of course, a robot could have the rights of a thing or a legal entity, if it could satisfy the relevant criteria.

V. Metaphors on Consciousness

"Imagine a world alive with incomprehensible objects and shimmering with an endless variety of movement and innumerable gradations of color. Imagine a world before the 'beginning was the word.'"[57]

Philosophers often use metaphors to illustrate their ideas, and philosophers of mind are no exception. Specifically, philosophers often use metaphors to explain their theories of things they do not actually understand. A cynic might say that metaphors often camouflage implausible analogies. In any case, one could surely write a metaphorical history of philosophy, or at the very least, a history of philosophical metaphors. I will reflect on three metaphors common to the philosophy of mind: the lion, the bat, and the thermostat.

A. The Lion

"Long, Long Ago People Used To Say / You're A Lion / You're A Lion, Mama."[58]

For better or worse, Ludwig Wittgenstein is notoriously quotable.[59] And one of his most frequently quoted remarks is the gnomic aphorism, "If a lion could speak, we could not understand him."[60]

Ironically, the lion and the aphorism are equally inscrutable. Taken literally, the aphorism is obviously false. Lions can already "speak." They communicate with each other, and try to communicate with us, with at least some success. When a lion roars, you get the message. And zoologists study the purpose and meaning of lion vocalizations.[61] If we can interpret the meaning of expressions made by mute lions, surely we could understand a speaking lion.[62]

But a literal reading of the aphorism surely misses its point. Wittgenstein studied epistemology, not zoology. He wasn't interested in lions, but in the relationship between language and knowledge.

---

[57] Stan Brakhage, Metaphors on Vision 1 (1963) (2017).
[58] Solomon Linda, Mbube ("Lion") (1939).
[59] Indeed, Wittgenstein's relentlessly quotable inscrutability has been the subject of considerable levity. See, e.g., Joshua Landy, Lost Fragments of the *Philosophical Investigations*, Arcade: Literature, the Humanities, and the World, Sept. 6, 2015 at http://arcade.stanford.edu/blogs/lost-fragments-philosophical-investigations ("706. Time for a deep-sounding metaphor. Language is a wave. We try to swim against it, but we are the water. (That'll get me into Bartlett's.)").
[60] Ludwig Wittgenstein, Philosophical Investigations 223 (G.E.M. Anscombe, trans. 1953, 1998) ("Wenn ein Löwe sprechen könnte, wir könnten ihn nicht verstehen.").
[61] See, e.g., Judith A. Rudnai, The Social Life of the Lion: A study of the behaviour of wild lions (Panthera leo massaica [Newmann]) in the Nairobi National Park, Kenya 55 (1973, 2012).
[62] See Constantine Sandis, Understanding the Lion For Real, in Knowledge, Language and Mind: Wittgenstein's Thought in Progress (A. Marques & N. Venturinha, eds. (forthcoming)).

Many scholars have responded to the aphorism by arguing that we certainly could understand a speaking lion, but it would no longer be a lion. For example, Daniel Dennett observed, “I think, on the contrary, that if a lion could talk, that lion would have a mind so different from the general run of lion minds, that although we could understand him just fine, we would learn little about ordinary lions from him.”[63] And Stephen Budiansky argued that the aphorism “begs the question: if a lion could talk, we probably could understand him. He just would not be a lion any more; or rather, his mind would no longer be a lion's mind.”[64]

But these responses also miss the point of the aphorism. Wittgenstein was not talking about lions or minds, but epistemology. Under his theory of language, meaning depends on shared experience. We can understand the meaning of expressions made by other people, because we share the experience of being people: “Shared human behaviour is the system of reference by means of which we interpret an unknown language.”[65] But we cannot understand the meaning of expressions made by non-human animals, because we cannot share their experiences. Of course, we can attribute meaning to expressions made by non-human animals, but we are the ultimate source of that meaning, not the animal.

Surely, Wittgenstein intended his aphorism to express the proposition that we could not understand the meaning of a statement made by a lion, because we cannot share the experience of being a lion. The problem is not that lions cannot speak, but that the experience of being a lion cannot be expressed in a form that we can understand. Or rather, we could not understand a speaking lion, because its statements would express an alien experience of the world.

And yet, given the run of the bestiary, why did Wittgenstein speak of a lion? Historically, the lion was a metaphor for Christ: As the lion is the king of beasts, so too is Christ the king of men.[66] Perhaps Wittgenstein’s aphorism was a veiled reference to Aslan, the speaking lion and soteriological counterpart of Christ in C.S. Lewis’s novel, *The Lion, The Witch and the Wardrobe,* which was published shortly before Wittgenstein’s death in 1951.[67] Notably, Wittgenstein was sympathetic to Catholicism, and expressed a conflicted belief in the divinity of Christ.[68]

---

[63] Daniel Dennett, Consciousness Explained 447 (1991).
[64] Stephen Budiansky, If a Lion Could Talk: Animal Intelligence and the Evolution of Consciousness (1998).
[65] Ludwig Wittgenstein, Philosophical Investigations s. 206.
[66] Benjamin Keach, Preaching From the Types and Metaphors of the Bible, “Christ a Lion,” at 353 (c. 1800).
[67] C.S. Lewis, The Lion, The Witch and the Wardrobe (1950). Interestingly, G. E. M. Anscombe, the translator and editor of Philosophical Investigations, was a devout Catholic who had debated Lewis on the truth of metaphysical naturalism.
[68] Ludwig Wittgenstein, Culture and Value 33 (1977). “What inclines even me to believe in Christ's resurrection? I play as it were with the thought.–If he did not rise from the dead, then he decomposed in the grave like every human being. He is dead & decomposed. In that case he is a teacher, like any other & can no longer help; & we are once more orphaned & alone. And have to make do with wisdom & speculation. It is as though we are in a hell, where we can only dream & are shut out from heaven, roofed

Lewis's lion was a metaphor for spiritual knowledge: When it speaks, we understand. Perhaps Wittgenstein's lion was a metaphor for agnosticism: When it speaks, we cannot understand. "Suppose someone said: 'What do you believe, Wittgenstein? Are you a sceptic? Do you know whether you will survive death?' I would really, this is a fact, say 'I can't say. I don't know', because I haven't any clear idea what I am saying when I am saying, 'I don't cease to exist.'"[69]

B. The Bat

"I didn't know that love would strike me / But this is what it's like / This is what it's like."[70]

In his seminal essay, "What is it like to be a bat?," Thomas Nagel used a variation on Wittgenstein's aphorism as a metaphor to question psychophysical reductionism.[71] Essentially, he argued that a physical theory of mind must provide a physical explanation of subjective experience.

Nagel assumed that at least some non-human animals are conscious, and consciousness implies subjective experience. Accordingly, conscious animals must have subjective experiences. "[T]he fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism."[72] So, if bats are conscious, then a bat must have the subjective experience of being a bat: "the essence of the belief that bats have experience is that there is something that it is like to be a bat."[73]

But the subjective experience of being a bat is impossible for a human to imagine. "Even without the benefit of philosophical reflection, anyone who has spent some time in an enclosed space with an excited bat knows what it is to encounter a fundamentally *alien* form of life."[74] For example, some bats use echolocation to locate and identify objects, emitting ultrasonic calls and interpreting the echoes to perceive the world around them. While bat echolocation is a form of perception analogous to human vision, the subjective experience of perceiving the world via echolocation is inconceivable for humans, who cannot perceive the world in that way.

---

in as it were. But if I am to be REALLY redeemed,– I need certainty–not wisdom, dreams, speculation–and this certainty is faith. And faith is faith in what my heart, my soul, needs, not my speculative intellect. For my soul, with its passions, as it were with its flesh and blood, must be redeemed, not my abstract mind. Perhaps one may say: Only love can believe the Resurrection. Or: it is love that believes the Resurrection."

[69] Ludwig Wittgenstein, Lectures and Conversations on Aesthetics, Psychology and Religious Belief 70 (2007).

[70] Arthur Russell, *What It's Like*, Love is Overtaking Me (2008).

[71] Thomas Nagel, What is it like to be a bat?, 83 Phil. Rev. 4 (1974).

[72] Thomas Nagel, What is it like to be a bat?, 83 Phil. Rev. 4 (1974).

[73] Thomas Nagel, What is it like to be a bat?, 83 Phil. Rev. 4 (1974).

[74] Thomas Nagel, What is it like to be a bat?, 83 Phil. Rev. 4 (1974) (emphasis in original).

In other words, we can understand what bats do, but we cannot understand what they experience. Of course, neither could a bat understand what humans experience. By contrast, while we cannot directly experience the subjective experiences of other humans, we can understand their subjective experiences, but only because we have similar subjective experiences. Facts about subjective experience can be objective only because humans share similar subjective experiences.

Nagel argues that this is a problem for psychophysical reductionism, because it cannot provide an objective account of subjective experience that does not itself ultimately rely on subjective experience. “If we acknowledge that a physical theory of mind must account for the subjective character of experience, we must admit that no presently available conception gives us a clue how this could be done.”[75]

Psychophysical reductionism holds that mental states are brain states and mental events are physical events. This may be true. Indeed, it probably is true. But we do not understand why. “At the present time the status of physicalism is similar to that which the hypothesis that matter is energy would have had if uttered by a pre-Socratic philosopher. We do not have the beginnings of a conception of how it might be true.”[76]

C. The Thermostat

“With most people, unbelief in one thing springs from blind belief in another.”[77]

1. Dennett’s Intentional Thermostat

In his book, *The Intentional Stance*, Daniel Dennett used a thermostat as a metaphor for an “intentional system.” He argued that a “system” has “beliefs” if its behavior is predicted by an “intentional strategy.” For example, a thermostat has beliefs because an intentional strategy can predict its behavior: “the thermostat will turn off the boiler as soon as it comes to believe the room has reached the desired temperature.”[78] As a consequence, Dennett insists that consciousness is nothing more than predictable behavior. “The perverse claim remains: *all there is* to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy, and hence *all there is* to really and truly believing that *p* (for an proposition *p*) is being an intentional system for which *p* occurs as a belief in the best (most predictive) interpretation.”[79]

Of course, even if a thermostat is an intentional system that can have beliefs and desires, it can’t have very many: “it can believe the room is too cold or too hot, that the boiler is on or off,

---

[75] Thomas Nagel, What is it like to be a bat?, 83 Phil. Rev. 4 (1974).
[76] Thomas Nagel, What is it like to be a bat?, 83 Phil. Rev. 4 (1974).
[77] Georg Christoph Lichtenberg, The Waste-Books, L81 (1778).
[78] Daniel C. Dennett, The Intentional Stance 22 (1987).
[79] Daniel C. Dennett, The Intentional Stance 29 (1987) (emphasis in original).

and that if it wants the room warmer it should turn on the boiler, and so forth."[80] But those beliefs and desires can be abstracted and multiplied. Eventually, a thermostat may come to have beliefs "about heat," not only because it regulates heat, but also because it can only regulate heat, and regulates heat in increasingly complicated ways.

According to Dennett, a "belief" or "internal representation" is just a feature of an intentional system. "There is no magic moment in the transition from a simple thermostat to a system that *really* has an internal representation of the world around it. The thermostat has a minimally demanding representation of the world, fancier thermostats have more demanding representations of the world, fancier robots for helping around the house would have still more demanding representations of the world. Finally you reach us."[81] In other words, a person is just a really complicated thermostat.

Of course, understanding how a thermostat works is easy, but understanding how a person works is hard. No matter. According to Dennett, it is only a matter of degree, albeit of great degree. "The principles, and problems, of interpretation that we discover when we attribute beliefs to people are the *same* principles and problems we discover when we look at the ludicrous, but blessedly simple, problem of attributing beliefs to a thermostat. The differences are of degree, but nevertheless of such great degree that understanding the internal organization of a simple intentional system gives one very little basis for understanding the internal organization of a complex intentional system, such as a human being."[82]

This increase in complexity is what distinguishes a thermostat from a person. "As we ascend the scale of complexity from simple thermostat, through sophisticated robot, to human being, we discover that our efforts to design systems with the requisite behavior increasingly run foul of the problem of *combinatorial explosion*. Increasing some parameter by, say, ten percent - ten percent more inputs or more degrees of freedom in the behavior to be controlled or more words to be recognized or whatever - tends to increase the internal complexity of the system being designed by orders of magnitude.[83] According to Dennett, "our brains will eventually be understood as symbol manipulating systems in at least rough analogy with computers." If brain matter is the hardware, then human language is the software, an "elegant, *generative*, indefinitely extendable" system of representation.[84]

2. Chalmers's Conscious Thermostat

In his book *The Conscious Mind*, David Chalmers turned Dennett's thermostat metaphor on its head, arguing that consciousness may be associated with any information-processing system,

---

[80] Daniel C. Dennett, The Intentional Stance 30 (1987) (emphasis in original).
[81] Daniel C. Dennett, The Intentional Stance 32 (1987) (emphasis in original).
[82] Daniel C. Dennett, The Intentional Stance 32-33 (1987) (emphasis in original).
[83] Daniel C. Dennett, The Intentional Stance 34-35 (1987) (emphasis in original).
[84] Daniel C. Dennett, The Intentional Stance 35 (1987) (emphasis in original).

including a simple system like a thermostat.[85] And he did so by repurposing Nagel's question and asking, "What is it like to be a thermostat?"

Chalmers adopted a dualist theory of mind, holding that consciousness is fundamental property of the world. And he argued that consciousness probably is associated with any information-processing system, no matter how simple. Humans are complex systems, with complex experiences. Mice are simpler systems, with simpler experiences, but presumably they are still conscious. "Mice may not have much of a sense of self, and may not be given to introspection, but it seems entirely plausible that there is *something* it is like to be a mouse."[86] A mouse has perceptions; surely it also has experiences.

But the same must be true of even simpler animals, like lizards, fishes, and slugs. They too have perceptions, and therefore must have experiences. So, when does experience disappear? If a slug, or plant, or tardigrade can perceive the world, and therefore experience the world, why not a thermostat. "The thermostat seems to realize the sort of information processing in a fish or a slug stripped down to its simplest form, so perhaps it might also have the corresponding sort of phenomenology in its most stripped-down form."[87]

What could a thermostat perceive? Not much. "Let us consider an information-processing system that is almost maximally simple: a thermostat. Considered as an information-processing device, a thermostat has just three information states (one state leads to cooling, another to heating, and another to no action). So the claim is that to each of these information states, there corresponds a phenomenal state. These three phenomenal states will all be different, and changing the information state will change the phenomenal state."[88]

But maybe it is enough to support some rudimentary form of consciousness? "Certainly it will not be very interesting to be a thermostat. The information processing is so simple that we should expect the corresponding phenomenal states to be equally simple. There will be three primitively different phenomenal states, with no further structure. . . . We will likely be unable to sympathetically imagine these experiences any better than a blind person can imagine sight, or than a human can imagine what it is like to be a bat; but we can at least intellectually know something about their basic structure."[89]

Of course, it is hard for us to imagine how a thermostat could have an experience. Where would the experience happen? But if consciousness is a fundamental property, then it can be associated with a thermostat in the same way that it is associated with a brain. "However we

---

[85] David J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory (1997). See also David J. Chalmers, *What is it like to be a Thermostat?,* at http://consc.net/notes/lloyd-comments.html (commenting on Dan Lloyd, *What is it Like to Be a Net?*).
[86] David J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory (1997).
[87] David J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory (1997).
[88] David J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory (1997).
[89] David J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory (1997).

make sense of this relation, the same will apply to thermostats: strictly speaking it is probably best not to say that the thermostat has the experiences (although I will continue to say this when talking loosely), but that the experiences are associated with the thermostat."[90]

Indeed, if consciousness is a fundamental property, we should expect it to be associated with every system, no matter how simple. "If there is experience associated with thermostats, there is probably experience everywhere: wherever there is a causal interaction, there is information, and wherever there is information, there is experience."[91] Effectively, Chalmers advanced a version of "panpsychism," or the idea that consciousness is everywhere, and everything in the universe is conscious.

VI. Metaphors in Action

"We do not think metaphors are very important, but a good metaphor is something even the police should watch."[92]

So, we have seen how philosophers use metaphor to explore intuitions about consciousness. But why these metaphors? And why metaphor at all? The rhetoric of consciousness is ineluctably metaphorical. We can no more escape metaphors on consciousness than we can escape metaphors on illness. As Susan Sontag observed, "[I]t is hardly possible to take up one's residence in the kingdom of the ill unprejudiced by the lurid metaphors with which it has been landscaped. It is toward an elucidation of those metaphors, and a liberation from them, that I dedicate this inquiry."[93] So, too, with consciousness, itself already a metaphor. But from whence do these metaphors come, and how shall we endeavor to liberate ourselves from them, or at least peer through them, and cast our gaze upon some fragment of that which they obscure?

A. Compulsive Anthropomorphism

A dog
that dies
and knows
that it dies
like a dog

and who can say
that it knows
that it dies
like a dog

---

[90] David J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory (1997).
[91] David J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory (1997).
[92] Georg Christoph Lichtenberg, The Waste-Books, E91 (1776).
[93] Susan Sontag, Illness as Metaphor 3-4 (1978).

is a man.[94]

Common sense and experience confirm that many non-human animals are conscious and have subjective experiences. Indeed, it is at least possible that all living things have some form of subjective experience and are conscious of the world in some way, even if we cannot possibly comprehend their subjective experience or consciousness. Not only mammals, but also reptiles and fish can be trained to solve problems. Insects engage in complicated social behaviors. The lowly flatworm makes choices. And even plants respond to external stimuli.

Perhaps consciousness requires some or all of the biological structures found in brains. Or perhaps a rudimentary form of consciousness - pre-consciousness? - is a consequence of the transition from mineral to animal. Perhaps even viruses are at the cusp of pre-consciousness, "organisms at the edge of life," which possess genes, evolve, and reproduce, but lack a cellular structure.[95] Of course, if it is impossible to imagine what it is like to be a bat, it is even more impossible to imagine what it is like to be a plant or a virus. But our imaginations are not the measure of the world.

Nevertheless, we constantly and unselfconsciously imagine what it is like to be a non-human animal. We just always imagine that the experience of being a non-human animal is exactly the same as the experience of being a human. Our anthropomorphism is compulsive, the very lens through which we view the world. As we create God in our own image, and place ourselves in the center of the phenomenology of the universe, so too do we create imaginary animals in our own image, strange people clothed in fur, feather, and scale, performing peculiar translations of human rituals. But even as our "compulsive anthropomorphism" presumes to dignify animals by enabling us to imagine them as "almost human," it precludes us from imagining their actual experiences or even understanding their actual efforts to communicate.[96] Lions can talk, but we cannot - we *will not* - understand them.

B. Digital Anthropomorphism

"So God created man in his *own* image, in the image of God created he him."[97]

Our compulsive anthropomorphism is hardly limited to animals. We anthropomorphize everything, including the machines we make in our own image. Children attribute personalities to their dolls, and adults attribute personalities to their vehicles. Unsurprisingly, we are inclined to imagine the computer, a machine designed to reproduce the form of human thought, as a

---

[94] Erich Fried, Definition (1964).
[95] E.P. Rybicki, The classification of organisms at the edge of life, or problems with virus systematics, 86 South African Journal of Science 182 (1990).
[96] Stephen Budiansky, If a Lion Could Talk: Animal Intelligence and the Evolution of Consciousness xvii (1998).
[97] King James Bible, Genesis 1:27 (emphasis in original).

potential candidate for personhood. Indeed, we are inclined to imagine that because we created computers in our own image, as “thinking machines” capable of feats of logical reasoning we can dream but cannot accomplish, they replicate how we think.

The speculative literature on strong artificial intelligence - of course there is no other kind - is replete with apocalyptic warnings of superintelligent computers that will swiftly render us obsolete.[98] Our pharaohs of science and technology imagine themselves brought low by their own creations and tremble. “And children shall rise up against *their* parents, and shall cause them to be put to death.”[99]

But there is no particular reason to believe these fears are anything but Oedipal fantasies. A computer is a Turing machine, and computers asymptotically approach universal Turing machines, but brains are not necessarily Turing machines, and human brains are not necessarily universal Turing machines. Indeed, there is every reason to assume they are not. While we have learned many things about brains, we do not actually know how brains work. But we do know that brains do not appear to work like computers, as much as we like to use computers as analogies for brains, and vice versa.

A brain is a collection of neurons, axons, synapses, and perhaps other as yet unidentified or misunderstood structures. Some brains are larger and more complicated than others. A human brain has about 86 billion neurons and 150 trillion synapses, a fruit fly brain has about 250,000 neurons and 10 million synapses, and a *Caenorhabditis elegans* roundworm brain has 302 neurons and about 7,500 synapses. While scientists have mapped every neuron in the *C. Elegans* brain, one of the simplest nervous systems of any animal, they do not know how it works. And they don’t have the first clue how the human brain works, or how brains cause consciousness. It remains the domain of philosophical speculation, not science.

By contrast, the central processing unit (“CPU”) or “brain” of a digital computer is a collection of circuits that manipulate and store data in some medium. The earliest computers had a relatively small number of circuits, but the a modern computer may have billions or trillions of circuits. As the structure of computers became increasingly complicated, it became increasingly difficult for humans to design every circuit in a computer, and engineers had to use computers in order to design new computers. But in principle, humans can understand the purpose and function of every element of a computer. We know how computers work.

[98] See, e.g., Nick Bostrom, Superintelligence: Path, Dangers, Strategies (2014).
[99] King James Bible, Mark 13:12.

Computers are designed to reproduce human reasoning, or how we think about thinking.[100] They are not and cannot be designed to reproduce how we actually think, because we don't know how we think. "Language may well be *the* thing that makes possible the all-important leap from merely having intentions and beliefs to having intentions and beliefs about intentions and beliefs.[101] We know that thinking machines are possible, because brains are thinking machines. [102] But we do not know what kinds of machines can be thinking machines, because we don't know how thinking works.

The computational theory of mind assumes that consciousness, however defined, can be realized in any physical form, including a digital computer. Of course, it is conceivable that computers can replicate brains. But it is equally conceivable that they cannot. At the very least, we know that while computers reproduce a useful facsimile of human reasoning, they do not - and are not intended to - reproduce human thinking. Indeed, one of the reasons that computers are so useful is that the way they function enables them to do things that human brains cannot do well, if at all. Perhaps the flip-side of this remarkable functionality is that computers cannot do things that humans do well, like "thinking," however defined.

VII. Computer Intelligence as Tertium Quid

"As if beyond will or fate he and his beasts and his trappings moved both in card and in substance under consignment to some third and other destiny."[103]

When we speak of "artificial intelligence," we almost always mean "computer intelligence." Or rather, we mean the realization of consciousness or its functional equivalent on a computer. Both apostles and skeptics of artificial intelligence assume that proof of concept depends on whether a computer can replicate a brain. After all, if a computer can replicate the functionality of even the simplest brains - say the 302 neurons of *C. Elegans* - then it is only a matter of scale for it to represent the functionality of any brain.

But what if a computer cannot replicate the functionality of a brain? Recall, we do not actually know how brains work. It is at least conceivable that one or more physical elements critical to the functionality of a brain cannot be replicated in a computer. Perhaps a computer cannot

---

[100] Stephen Budiansky, If a Lion Could Talk: Animal Intelligence and the Evolution of Consciousness (1998) ("The basic wiring of all brains, as the early artificial intelligence researchers found to their great frustration, did not conform to that of a general-purpose digital computer that manipulates data in sequential, logical operations. But a brain plus language does - if it is admittedly a slow and rather plodding general-purpose problem solver.").
[101] Stephen Budiansky, If a Lion Could Talk: Animal Intelligence and the Evolution of Consciousness (1998).
[102] See generally John R. Searle, The Mystery of Consciousness (1997) and John R. Searle, Mind: A Brief Introduction (2004).
[103] Cormac McCarthy, Blood Meridian 96 (1985). See generally Michael Lynn Crews, Books Are Made of Books: A Guide to Cormac McCarthy's Literary Influences (2017).

actually replicate the functionality of a neuron or synapse or some other element of a brain, for some reason we have not yet identified and do not yet understand.

Would that necessarily mean a computer cannot be conscious? It depends on our definition of "consciousness." Certainly, it presumably would mean that computers could not be conscious in the way that humans are conscious. But then, it is already the case that nothing but a human can be conscious in the way that human can be conscious, just as nothing but a bat can be conscious in the way that a bat is conscious, and nothing but a roundworm can be conscious in the way that a roundworm is conscious, if it is conscious at all.

The question is whether it would be possible for a computer to be conscious in any way, not just in the way that an animal with a brain is conscious. After all, the overwhelming majority of living things lack brains. Plants lack brains, but react to external stimuli. Single-celled animals lack brains, but appear to engage in intentional behavior. Not even all multicellular animals have "brains" Sponges and *Trichoplax adhaerens* have no brain or nervous system at all, and have neither neurons nor synapses. Cnidarians do not have brains or nervous systems, but do have neurons that form a "nerve net." And ctenophores do not have brains or nervous systems, but do have nerve cells that evolved separately from all other animals and have a different biochemistry.

But perhaps a computer could become conscious in a way that we cannot currently understand. Ctenophores prove that a nervous system, and maybe even a brain, could develop in more than one way. In some sense, they are our closest approximation of an alien species, as their nervous system is fundamentally different than the nervous system of any other animal. Presumably, a ctenophore's subjective experience, such as it is, is also fundamentally different from the subjective experience of any bilaterian animal.

If more than one biological structure can support a nervous system, perhaps a non-biological structure can also support a "nervous system," a "brain," or a form of "consciousness." But is there any reason to assume that we would recognize it as such? It would be a fundamentally "alien" form of consciousness, farther removed from our own than a roundworm, a ctenophore, a sponge, a single-celled organism, or even a virus. It would be a "conscious thermostat," no matter how complex, a form of "life" so far removed from our own that surely we would struggle to even recognize it as such.

And if such a form of consciousness is possible, how would we know if and when it came into existence? Perhaps it already exists, as ignorant of us as we are ignorant of it, existing as it were in parallel planes of consciousness, experiencing mutually inconceivable worlds and thinking mutually incomprehensible thoughts. Would it not be the ultimate irony if we were to unwittingly become the God we imagined and create "Machine," fashioned in our own image, but utterly alien to ourselves, dreaming of a world we cannot imagine or ever inhabit?