Legal Personhood for Artificial Intelligences

Lawrence B. Solum

ESSAY

LEGAL PERSONHOOD FOR ARTIFICIAL INTELLIGENCES

LAWRENCE B. SOLUM*

Could an artificial intelligence become a legal person? As of today, this question is only theoretical. No existing computer program currently possesses the sort of capacities that would justify serious judicial inquiry into the question of legal personhood. The question is nonetheless of some interest. Cognitive science begins with the assumption that the nature of human intelligence is computational, and therefore, that the human mind can, in principle, be modelled as a program that runs on a computer.[1] Artificial intelligence (AI) research attempts to develop such models.[2] But even as cognitive science has displaced behavioralism as

* Professor of Law and William M. Rains Fellow, Loyola Law School, Loyola Marymount University. B.A. 1981, University of California at Los Angeles; J.D. 1984, Harvard Law School. I owe thanks to Scott Altman, Ken Anderson, Don Brosnan, Don Crenshaw, Zlatan Damnjanovic, Michael Fitts, Kent Greenawalt, Sharon Lloyd, Shelley Marks, David Millon, Elyn Saks, and Paul Weithman for comments made on earlier versions of this essay. My colleagues Dave Leonard, Sam Pillsbury, Dave Tunick, and Peter Tiersma have been generous in sharing criticisms and suggestions. Bill Mulherin of the William M. Rains Law Library and Jai Gohel of the Loyola Law School Class of 1992 provided valuable research assistance. Finally, I am grateful to the editors of this review for their many helpful suggestions.

1. For an introduction to cognitive science and the philosophy of mind, see OWEN J. FLANAGAN, JR., THE SCIENCE OF THE MIND 1-22 (2d ed. 1991). For the purposes of this essay, I will not address the question as to which computer architectures could produce artificial intelligence. For example I will not discuss the question whether parallel, as opposed to serial, processing would be required. Similarly, I will not discuss the merits of connectionist as opposed to traditional approaches to AI. For a comparison of parallel distributed processing with serial processing, see *id.* at 224-41. These issues are moot in one sense. A digital computer can, in principle, implement any connectionist or parallel approach. On the other hand, there could be one very important practical difference: the parallel architecture could turn out to be much faster.

2. There is a debate within the artificial intelligence community as to the goal of AI research. The possibilities range from simply making machines smarter to investigating the nature of human intelligence or, more broadly, the nature of all intelligence. *See* Bob Ryan, *AI's Identity Crisis*, BYTE, Jan. 1991, at 239, 239-40. Owen Flanagan distinguishes four programs of AI research. Nonpsychological AI research involves building and programming computers to accomplish tasks that would require intelligence if undertaken by humans. Weak psychological AI views computer models as a tool for investigating human intelligence. Strong psychological AI assumes that human minds really are computers and therefore in principle can be duplicated by AI research. Suprapsychological AI investigates the nature of all intelligence and hence is not limited to investigating the human mind. *See* FLANAGAN,

the dominant paradigm for investigating the human mind, fundamental questions about the very possibility of artificial intelligence continue to be debated. This Essay explores those questions through a series of thought experiments that transform the theoretical question whether artificial intelligence is possible into legal questions such as, "Could an artificial intelligence serve as a trustee?"

What is the relevance of these legal thought experiments for the debate over the possibility of artificial intelligence? A preliminary answer to this question has two parts. First, putting the AI debate in a concrete legal context acts as a pragmatic Occam's razor. By reexamining positions taken in cognitive science or the philosophy of artificial intelligence as legal arguments, we are forced to see them anew in a relentlessly pragmatic context.[3] Philosophical claims that no program running on a digital computer could really be intelligent are put into a context that requires us to take a hard look at just what practical importance the missing reality could have for the way we speak and conduct our affairs. In other words, the legal context provides a way to ask for the "cash value" of the arguments. The hypothesis developed in this Essay is that only some of the claims made in the debate over the possibility of AI do make a pragmatic difference, and it is pragmatic differences that ought to be decisive.[4]

Second, and more controversially, we can view the legal system as a repository of knowledge—a formal accumulation of practical judgments.[5] The law embodies core insights about the way the world works

---

supra note 1, at 241-42. This Essay discusses the philosophical foundations for Flanagan's third and fourth categories of AI research.

3. *See generally* Catharine W. Hantzis, *Legal Innovation Within the Wider Intellectual Tradition: The Pragmatism of Oliver Wendell Holmes, Jr.*, 82 NW. U. L. REV. 541, 561-75, 595-99 (1988) (discussing Holmes's jurisprudential focus on concrete issues rather than generalities); Steven D. Smith, *The Pursuit of Pragmatism*, 100 YALE L.J. 409, 409-12 (1990) (discussing the renewed popularity of legal pragmatism); *Symposium: The Renaissance of Pragmatism in American Legal Thought*, 63 S. CAL. L. REV. 1569 (1990) (collecting articles espousing diverse views of legal pragmatism). The law is a "pragmatic" context in the sense that legal decisions are made for practical purposes with consequences in mind and in the sense that foundationalist philosophical theories do not play a role in legal reasoning. This assertion does not depend on the further claim that legal actors have adopted American pragmatism as part of their world view.

4. In addition, the Essay advances a more modest hypothesis. Examining the debate over the possibility of AI through legal examples illuminates the consequences of the arguments made in the debate, and this pragmatic assessment has a bearing on the arguments, even if it is not decisive. This Essay surely does not suffice to confirm the ambitious hypothesis in the text. I hope that it has established the more modest claim presented in this footnote. The proof, of course, is in the pudding.

5. The point is related to that made by Sir John Fortescue about the common law of England. There is a presumption in favor of its wisdom, because it has been tested by long experience. *See* SIR JOHN FORTESCUE, DE LAUDIBUS LEGUM ANGLIE 39-41 (S.B. Chrimes

and how we evaluate it. Moreover, in common-law systems judges strive to decide particular cases in a way that best fits the legal landscape—the prior cases, the statutory law, and the constitution.[6] Hence, transforming the abstract debate over the possibility of AI into an imagined hard case forces us to check our intuitions and arguments against the assumptions that underlie social decisions made in many other contexts. By using a thought experiment that explicitly focuses on wide coherence,[7] we increase the chance that the positions we eventually adopt will be in reflective equilibrium[8] with our views about related matters. In addition, the law embodies practical knowledge in a form that is subject to public examination and discussion. Legal materials are published and subject to widespread public scrutiny and discussion. Some of the insights gleaned in the law may clarify our approach to the artificial intelligence debate.[9]

I do not claim in this Essay to have resolved the debate over the possibility of artificial intelligence. My aim is more modest: I am proposing a way of thinking about the debate that just might result in progress. There is some precedent for this project. Christopher Stone brought questions of environmental ethics into focus by asking whether trees should have standing.[10] My hope is that the law will be equally

---

ed., Wm. W. Gaunt & Sons 1986) (1537); *see also* J.G.A. POCOCK, THE MACHIAVELLIAN MOMENT: FLORENTINE POLITICAL THOUGHT AND THE ATLANTIC REPUBLICAN TRADITION 9-18 (1975) (offering a critique of Fortescue's laudatory expositions of English law and arguing that because English law is based on accumulated experiences—elevated to the level of custom—its very existence presumes its validity, thus preempting rational scrutiny of English law's assumption that it is well suited to the needs of the English).

6. Here I adopt the view of Ronald Dworkin. *See* RONALD DWORKIN, LAW'S EMPIRE 91-96, 147-50, 276-400 (1986); RONALD DWORKIN, A MATTER OF PRINCIPLE 3-42 (1985); RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 1-80 (1977).

7. *Cf.* SUSAN L. HURLEY, NATURAL REASONS 15-18, 225-53 (1987) (advocating coherence approach to reasoning in general and legal reasoning in particular); S.L. Hurley, *Coherence, Hypothetical Cases, and Precedent*, 10 OXFORD J. LEGAL STUD. 221, 222-32 (1990) (same, with emphasis on legal reasoning).

8. *See* JOHN RAWLS, A THEORY OF JUSTICE 48-51 (1971).

9. I say "may provide" advisedly. We must be on guard against an easy or unthinking move from a legal conclusion to a moral one. In many circumstances, there are good reasons for answering a moral question differently from a legal one. Most obviously, the costs of legal enforcement of a norm are quite different than the costs of social enforcement of a moral norm. Moreover, in many cases, the law will simply be morally wrong. The fact that a legal rule has survived a very long time does tell us that it has not led to the collapse of the society that enforces it, but it does not tell us directly whether that society would be better off without it. The legal case may bear on the moral one, but not being irrelevant is far from being decisive. I owe thanks to Elyn Saks for prompting me to qualify my argument in this regard.

10. *See* Christopher Stone, *Should Trees Have Standing?—Toward Legal Rights for Natural Objects*, 45 S. CAL. L. REV. 450, 453-57 (1972) [hereinafter Stone, *Should Trees Have Standing?*]. So far as I know, Stone became the first legal thinker to raise the questions asked by this Essay in a footnote to his famous 1972 essay. *See id.* at 456 n.26 (raising the question as to whether analysis applicable to natural objects such as trees would be appropriate to

fruitful as a context in which to think about the possibility of AI. The "artificial reason and judgment of law"[11] may circumvent the intractable intuitions that threaten to lock the AI debate in dialectical impasse.

Part I of this Essay recounts some recent developments in cognitive science and explores the debate as to whether artificial intelligence is possible. Part II puts the question in legal perspective by setting out the notion of legal personhood. Parts III and IV explore two hypothetical scenarios. Part III examines the first scenario—an attempt to appoint an AI as a trustee. The second scenario, an AI's invocation of the individual rights provisions of the United States Constitution, is the subject of Part IV. The results are then brought to bear on the debate over the possibility of artificial intelligence in Part V. In conclusion, Part VI takes up the question whether cognitive science might have implications for current legal and moral debates over the meaning of personhood.

I. ARTIFICIAL INTELLIGENCE

Is artificial intelligence possible? The debate over this question has its roots at the very beginning of modern thought about the nature of the human mind. It was Thomas Hobbes who first proposed a computational theory of mind: "By ratiocination, I mean computation."[12] And it was Rene Descartes who first considered a version of the question whether it would be possible for a machine to think:

> For we can easily understand a machine's being constituted so that it can utter words, and even emit some responses to action on it of a corporeal kind, which brings about a change in its organs; for example, if it is touched in a particular part it may ask what we wish to say to it; if in another part it may exclaim that it is being hurt, and so on, but it never happens that it arranges its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do.[13]

"computers"). Stone returned to this issue in 1987. *See* CHRISTOPHER STONE, EARTH AND OTHER ETHICS 12, 28-30, 65-67 (1987) [hereinafter STONE, EARTH AND OTHER ETHICS] (asking whether a "robot" should have standing and discussing criminal liability of AIs).

11. The phrase is Sir Edward Coke's. *See* Prohibitions Del Roy, 12 Coke Rep. 63, 65, 77 Eng. Rep. 1342, 1343 (1608).

12. THOMAS HOBBES, ELEMENTS OF PHILOSOPHY (1655), *reprinted in* 1 THE ENGLISH WORKS OF THOMAS HOBBES 1, 3 (William Molesworth ed., London, J. Bohn 1839); *see also* THOMAS HOBBES, LEVIATHAN (1670), *reprinted in* 3 THE ENGLISH WORKS OF THOMAS HOBBES 1, *supra*, at 29-32 [hereinafter LEVIATHAN] (equating reason with computation or "reckoning of the consequences"). Hobbes uses "ratiocination" to mean reasoning.

13. RENE DESCARTES, DISCOURSE ON THE METHOD OF RIGHTLY CONDUCTING ONE'S REASON AND SEEKING TRUTH IN THE SCIENCES (1637), *reprinted in* THE ESSENTIAL DESCARTES 138 (Margaret D. Wilson ed., 1969). This passage was likely inspired by

Descartes' assertion that no artifact could arrange its words "to reply appropriately to everything that may be said in its presence" remains at the heart of the AI debate.

The events of the past forty years have stretched the limits of our imagination. Digital computers have been programmed to perform an ever wider variety of complex tasks. As I write this Essay using a word processing program, my spelling and grammar are automatically checked by programs that perform tasks thought to require human intelligence not so many years ago. The program Deep Thought has given the second best human chess player a very tough game, and the program's authors predict the program will become the world's chess champion within a few years.[14] Expert systems simulate the thinking of human experts on a wide variety of subjects, from petroleum geology to law.[15]

But these events have not resolved the question whether AI is even possible. The contemporary debate[16] over that question has centered around Alan Turing's test.[17] Turing proposed that the question whether a machine can think be replaced with the following, more operationalized, inquiry. The artifact that is a candidate for having the ability to think shall engage in a game of imitation with a human opponent. Both

---

Descartes' experience with the French Royal Gardens, which included a miniature society inhabited by hydraulically animated robots. As visitors walked along garden paths they set the robots bodies in motion. The robots actually played musical instruments and spoke. *See* FLANAGAN, *supra* note 1, at 1-2.

14. *See* Feng-hsiung Hsu et al., *A Grandmaster Chess Machine*, SCI. AM., Oct. 1990, at 44, 44.

15. *See, e.g.*, ANN VON DER LIETH GARDNER, AN ARTIFICIAL INTELLIGENCE APPROACH TO LEGAL REASONING 1-24 (1987); RICHARD E. SUSSKIND, EXPERT SYSTEMS IN LAW: A JURISPRUDENTIAL INQUIRY (1987); ALAN TYREE, EXPERT SYSTEMS IN LAW 7-11 (1989); L. Thorne McCarty, *Artificial Intelligence and Law: How to Get There From Here*, 3 RATIO JURIS 189, 189-200 (1990); Edwina L. Rissland, *Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning*, 99 YALE L.J. 1957, 1961-64 (1990).

16. Overviews of the debate over the possibility of AI are found in JAMES H. FETZER, ARTIFICIAL INTELLIGENCE: ITS SCOPE AND LIMITS 3-27, 298-303 (1990); JOHN HAUGELAND, ARTIFICIAL INTELLIGENCE: THE VERY IDEA 2-12 (1985); RAYMOND KURZWEIL, THE AGE OF INTELLIGENT MACHINES 36-40 (1990); and in the essays collected in THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE (Margaret A. Boden ed., 1990) and in THE ARTIFICIAL INTELLIGENCE DEBATE: FALSE STARTS, REAL FOUNDATIONS (Stephen Graubard ed., 1988). For a strong statement of the view that AI is impossible, see HUBERT DREYFUS, WHAT COMPUTERS CAN'T DO: THE LIMITS OF ARTIFICIAL INTELLIGENCE 285-305 (rev. ed. 1979).

17. *See* Alan M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433 (1950), *reprinted in* THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE, *supra* note 16, at 40 (subsequent citations to pagination in anthology). For a recent discussion and defense of the Turing Test, see Daniel C. Dennett, *Can Machines Think?*, *in* KURZWEIL, *supra* note 16, at 48. For a recent critique, see Donald Davidson, *Turing's Test*, *in* MODELLING THE MIND 1 (K.A. Mohyeldin Said et al. eds., 1990). For a report on a recent competition testing present-day computers and programs in the Turing format, see Carl Zimmer, *Flake of Silicon*, DISCOVER, Mar. 1992, at 36, 36-38.

the candidate and the human being are questioned by someone who does not know which is which (or who is who)—the questions are asked via teletype. The questions may be on any subject whatsoever. Both the human being and the artifact will attempt to convince the questioner that it or she is the human and the other is not. After a round of play is completed, the questioner guesses which of the two players is the human. Turing suggested we postpone a direct answer to the question whether machines can think; he proposed that we ask instead whether an artifact could fool a series of questioners as often as the human was able to convince them of the truth, about half the time.[18] The advantage of Turing's test is that it avoids direct confrontation with the difficult questions about what "thinking" or "intelligence" is. Turing thought that he had devised a test that was so difficult that anything that could pass the test would necessarily qualify as intelligent.

John Searle questioned the relevance of Turing's Test with another thought experiment, which has come to be known as the Chinese Room.[19] Imagine that you are locked in a room. Into the room come batches of Chinese writing, but you don't know any Chinese. You are, however, given a rule book, written in English, in which you can look up the bits of Chinese, by their shape. The rule book gives you a procedure for producing strings of Chinese characters that you send out of the room. Those outside the room are playing some version of Turing's game. They are convinced that whatever is in the room understands Chinese. But you don't know a word of Chinese, you are simply following a set of instructions (which we can call a program) based on the shape of Chinese symbols. Searle believes that this thought experiment demonstrates that neither you nor the instruction book (the program) understands Chinese, even though you and the program can simulate such understanding.[20]

More generally, Searle argues that thinking cannot be attributed to a computer on the basis of its running a program that manipulates symbols in a way that simulates human intelligence. The formal symbol-manipulations accomplished by the program cannot constitute thinking or un-

18. *See* Turing, *supra* note 17, at 41-48.

19. *See* JOHN R. SEARLE, MINDS, BRAINS AND SCIENCE 28-41 (1984); John R. Searle, *Author's Response*, 3 BEHAVIORAL & BRAIN SCI. 450 (1980); John R. Searle, *Is the Brain a Digital Computer?*, 64 PROC. & ADDRESSES AM. PHIL. ASS'N, Nov. 1990, at 21, 21; John R. Searle, *Minds, Brains & Programs*, 3 BEHAVIORAL & BRAIN SCI. 417 (1980), *reprinted in* THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE, *supra* note 16, at 67 [hereinafter Searle, *Minds, Brains & Programs*; subsequent citations to pagination in anthology]; John R. Searle, *"The Emperor's New Mind": An Exchange*, N.Y. REV. BOOKS, June 14, 1990, at 58 (letter to the editor with response from John Maynard Smith).

20. Searle, *Minds, Brains & Programs*, *supra* note 19, at 70.

derstanding because the program lacks "intentionality"—the ability to process meanings. The shape of a symbol is a syntactic property, whereas the meaning of a symbol is a semantic property. Searle's point is that computer programs respond only to the syntactic properties of symbols on which they operate.[21]

This point can be restated in terms of the Chinese Room: (1) the output—coherent Chinese sentences—from the Chinese room seems to respond to the meaning of the input; (2) but the process that goes on inside the Chinese room only involves the shape or syntactic properties of the input; (3) therefore, the process in the Chinese room does not involve understanding.[22] Searle generalizes the conclusion of the Chinese room thought experiment by arguing that part of the definition of a program is that it is formal and operates only on syntactic properties. He concludes that no system could be said to think or understand solely on the basis of the fact that the system is running a program that produces output that simulates understanding.[23]

Searle's Chinese Room has given rise to a number of replies.[24] But

---

21. *Id.*

22. *Id.* at 83-84.

23. *Id.* at 70-71.

24. *See, e.g.*, Robert P. Abelson, *Searle's Argument Is Just a Set of Chinese Symbols*, 3 BEHAVIORAL & BRAIN SCI. 424 (1980); Ned Block, *What Intuitions About Homunculi Don't Show*, 3 BEHAVIORAL & BRAIN SCI. 425 (1980); Bruce Bridgeman, *Brains + Programs = Minds*, 3 BEHAVIORAL & BRAIN SCI. 427 (1980); Arthur C. Danto, *The Use and Mention of Terms and the Simulation of Linguistic Understanding*, 3 BEHAVIORAL & BRAIN SCI. 428 (1980); Daniel Dennett, *The Milk of Human Intentionality*, 3 BEHAVIORAL & BRAIN SCI. 428 (1980); John C. Eccles, *A Dualist-Interactionist Perspective*, 3 BEHAVIORAL & BRAIN SCI. 430 (1980); J.A. Fodor, *Searle on What Only Brains Can Do*, 3 BEHAVIORAL & BRAIN SCI. 431 (1980); John Haugeland, *Programs, Causal Powers, and Intentionality*, 3 BEHAVIORAL & BRAIN SCI. 432 (1980); Douglas R. Hofstadter, *Reductionism and Religion*, 3 BEHAVIORAL & BRAIN SCI. 433 (1980); B. Libet, *Mental Phenomena and Behavior*, 3 BEHAVIORAL & BRAIN SCI. 434 (1980); William G. Lycan, *The Functionalist Reply (Ohio State)*, 3 BEHAVIORAL & BRAIN SCI. 434 (1980); John C. Marshall, *Artificial Intelligence—The Real Thing?*, 3 BEHAVIORAL & BRAIN SCI. 435 (1980); Grover Maxwell, *Intentionality: Hardware, Not Software*, 3 BEHAVIORAL & BRAIN SCI. 437 (1980); John McCarthy, *Beliefs, Machines, and Theories*, 3 BEHAVIORAL & BRAIN SCI. 435 (1980); E.W. Menzel, Jr., *Is the Pen Mightier than the Computer?*, 3 BEHAVIORAL & BRAIN SCI. 438 (1980); Marvin Minsky, *Decentralized Minds*, 3 BEHAVIORAL & BRAIN SCI. 439 (1980); Thomas Natsoulas, *The Primary Source of Intentionality*, 3 BEHAVIORAL & BRAIN SCI. 440 (1980); Roland Puccetti, *The Chess Room: Further Demythologizing of Strong AI*, 3 BEHAVIORAL & BRAIN SCI. 441 (1980); Zenon W. Pylyshyn, *The "Causal Power" of Machines*, 3 BEHAVIORAL & BRAIN SCI. 442 (1980); Howard Rachlin, *The Behavioralist Reply (Stony Brook)*, 3 BEHAVIORAL & BRAIN SCI. 444 (1980); Martin Ringle, *Mysticism as a Philosophy of Artificial Intelligence*, 3 BEHAVIORAL & BRAIN SCI. 444 (1980); Richard Rorty, *Searle and the Special Powers of the Brain*, 3 BEHAVIORAL & BRAIN SCI. 445 (1980); Roger C. Shank, *Understanding Searle*, 3 BEHAVIORAL & BRAIN SCI. 446 (1980); Aaron Sloman & Monica Croucher, *How to Turn an Information Processor Into an Understander*, 3 BEHAVIORAL & BRAIN SCI. 447 (1980); William E. Smythe, *Simulation Games*, 3 BEHAVIORAL & BRAIN SCI. 448 (1980); Donald O. Walter, *The Thermostat and the*

at this point I will leave the debate over the possibility of AI.

First Interlude[25]

When Mike was installed in Luna, he was pure thinkum, a flexible logic—"High-Optional, Logical, Multi-Evaluating Supervisor, Mark IV, Mod. L"—a HOLMES FOUR. He computed ballistics for pilotless freighters and controlled their catapult. This kept him busy less than one percent of time and Luna Authority never believed in idle hands. They kept hooking hardware into him—decision-action boxes to let him boss other computers, bank on bank of additional memories, more banks of associational neural nets, another tubful of twelve-digit random numbers, a greatly augmented temporary memory. Human brain has around ten-to-the-tenth neurons. By third year Mike had better than one and half times that number of neuristors.

And woke up.

Am not going to argue whether a machine can "really" be alive, "really" be self-aware. Is a virus self-aware? Nyet. How about oyster? I doubt it. A cat? Almost certainly. A human? Don't know about you, tovarishch, but I *am. Somewhere along the evolutionary chain from macromolecule to human brain awareness crept in. Psychologists assert it happens automatically whenever a brain acquires certain very high number of associational paths. Can't see it matters whether paths are protein or platinum.*

—Robert A. Heinlein, *The Moon is a Harsh Mistress*

II. LEGAL PERSONHOOD

The classical discussion of the idea of legal personhood is found in John Chipman Gray's *The Nature and Sources of the Law*.[26] He began his famous discussion, "In books of the 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

---

Philosophy Professor, 3 BEHAVIORAL & BRAIN SCI. 449 (1980); Robert Wilensky, *Computers, Cognition and Philosophy*, 3 BEHAVIORAL & BRAIN SCI. 449 (1980).

25. ROBERT A. HEINLEIN, THE MOON IS A HARSH MISTRESS 13-14 (1966). Copyright 1966 by Robert A. Heinlein. Reprinted by permission of the Berkley Publishing Group.

26. *See* JOHN CHIPMAN GRAY, THE NATURE AND SOURCES OF THE LAW (Roland Gray ed., MacMillan 1921) (1909); *see also* Stephen C. Hicks, *On the Citizen and the Legal Person: Toward the Common Ground of Jurisprudence, Social Theory, and Comparative Law as the Premise of a Future Community, and the Role of the Self Therein*, 59 U. CIN. L. REV. 789, 808-21 (1991) (discussing the construct of the legal person in the context of social theory); Richard Tur, *The 'Person' in Law*, *in* PERSONS AND PERSONALITY: A CONTEMPORARY INQUIRY 116, 116-27 (Arthur Peacocke & Grant Gillett eds., 1987) (providing a concise summary of the concept of the person within several areas of the law).

duties."[27] The question whether an entity should be considered a legal person is reducible to other questions about whether or not the entity can and should be made the subject of a set of legal rights and duties.[28] The particular bundle of rights and duties that accompanies legal personhood varies with the nature of the entity. Both corporations and natural persons are legal persons, but they have different sets of legal rights and duties. Nonetheless, legal personhood is usually accompanied by the right to own property and the capacity to sue and be sued.

Gray reminds us that inanimate things have possessed legal rights at various times. Temples in Rome and church buildings in the middle ages were regarded as the subject of legal rights. Ancient Greek law and common law have even made objects the subject of legal duties.[29] In admiralty, a ship itself becomes the subject of a proceeding in rem and can be found "guilty."[30] Christopher Stone recently recounted a twentieth-century Indian case in which counsel was appointed by an appellate court to represent a family idol in a dispute over who should have custody of it.[31] The most familiar examples of legal persons that are not natural persons are business corporations and government entities.[32]

Gray's discussion was critical of the notion that an inanimate thing might be considered a legal person. After all, what is the point of making a thing—which can neither understand the law nor act on it—the subject of a legal duty?[33] Moreover, he argued that even corporations are reducible to relations between the persons who own stock in them, manage them, and so forth.[34] Thus, Gray insisted that calling a legal person a "person" involved a fiction unless the entity possessed "intelligence" and

---

27. GRAY, *supra* note 26, at 27.

28. This statement is not quite correct. As Christopher Stone points out, *X* may be given the legal status of personhood in order to confer rights on *Y*. Thus, giving a fetus the status of personhood might confer the right to sue in tort for injury to it on its parents. *See* STONE, EARTH AND OTHER ETHICS, *supra* note 10, at 43.

29. GRAY, *supra* note 26, at 46.

30. *See id.* at 48-49; Stone, *Should Trees Have Standing?*, *supra* note 10, at 5.

31. *See* STONE, EARTH AND OTHER ETHICS, *supra* note 10, at 22 (citing Mullick v. Mullick, 52 I.A. 245, 256-61 (P.C. 1925)).

32. *See* David Millon, *Theories of the Corporation*, 1990 DUKE L.J. 201, 206 (discussing historical development of the theory of corporations).

33. GRAY, *supra* note 26, at 48-49, 53. The corresponding question about the point of making an inanimate thing the subject of a legal right is easier to answer. Giving temples or trees rights that can be enforced by guardians or private attorneys general has an obvious objective—to protect the tree or temple. *See generally* Stone, *Should Trees Have Standing?*, *supra* note 10 (discussing implications of legal rights for the environment). Of course, the same sort of argument can be made for making inanimate objects the subjects of legal duties. The tree can be made liable for damage done by a falling branch to induce a natural person to take preventative action—calling the tree trimmers.

34. GRAY, *supra* note 26, at 50-51.

"will."[35] Those attributes are part of what is in contention in the debate over the possibility of AI.[36]

III. COULD AN ARTIFICIAL INTELLIGENCE SERVE AS A TRUSTEE?

This case study and the one that follows are intended to illustrate two different sorts of issues in the AI debate. In this first scenario, we explore the issue of competence (of "intelligence" in the sense of capacity to perform complex actions) by posing the question whether an AI could serve as a trustee. The second scenario explores the questions of intentionality and consciousness (of "will" in a sense) by asking whether an AI could claim the more robust rights of legal and moral personhood guaranteed by the Bill of Rights and the Civil War Amendments to the United States Constitution.

A. The Scenario

This first scenario speculates about the legal consequences of developing an expert system capable of doing the things a human trustee can do.[37] Imagine such expert systems developing from existing programs that perform some of the component functions of a trustee. For example, the decision to invest in publicly traded stocks is made by a computer program in what is called "program trading," in which the program makes buy or sell decisions based on market conditions.[38] Today, one

35. *Id.* at 52.

36. It is important to remember that the question whether something should be given legal personhood is distinct from the question whether it has moral rights. (I use the term "moral right" to refer to moral claim rights, that is, to moral claims that individuals have on one another, with "moral" used in contrast with "legal.") Thus, the fact that corporations are legal persons with constitutional rights—such as the rights to freedom of speech, due process, and equal protection of the laws—does not entail the conclusion that corporations have equivalent moral rights. Vice versa, the possession of moral rights does not lead automatically to the conclusion that there should be corresponding legal rights. *See* STONE, EARTH AND OTHER ETHICS, *supra* note 10, at 43, 73. This point is a narrow one. The factors that bear on the decision to grant legal rights may bear on the question whether corresponding moral rights exist, but the relationship between the two sorts of rights is not one of entailment in either direction.

37. For those unfamiliar with the common-law term "trust," it is defined as "a fiduciary relationship with respect to property, subjecting the person by whom the title to property is held to equitable duties to deal with the property for the benefit of another person, which arises as a result of a manifestation of an intention to create it." RESTATEMENT (SECOND) OF TRUSTS § 2 (1959). The trustee is the legal person who administers the trust—invests trust assets, and so forth. The beneficiary is the person for whom the trust is maintained, for example, the person who receives income from the trust. The settlor is the person who establishes the trust. The terms of a trust are the directives to the trustee in the document or instrument creating the trust.

38. *See* Christina Toh-Pantin, *Wall Street Sees Tide Turing on Program Trading*, Reuters

also can buy a computer program that will automatically issue instructions to pay your regular monthly bills by sending data to a bank or service via modem. It is not difficult to imagine an expert system that combines these functions with a variety of others, in order to automate the tasks performed by the human trustee of a simple trust.

Such a system might evolve in three stages. At stage one, the program aids a human trustee in the administration of a large number of simple trusts. The program invests in publicly traded securities, placing investment orders via modem and electronic mail. The program disburses the funds to the trust beneficiaries via an electronic checking program. Upon being informed of a relevant event, such as the death of a beneficiary, the program follows the instructions of the trust instrument—for example, changing the beneficiary or terminating the trust. The program prepares and electronically files a tax return for the trust. The human trustee operates as do trustees today. The human makes the ultimate decisions on how to invest the funds, although she may rely upon an expert system for advice. She reviews the program's activities to insure that the terms of the trust instrument are satisfied. But the actual performance of the day-to-day tasks is largely automated, carried out by the program without the need of human intervention.

Stage two involves a greater role for the AI. Expert systems are developed that outperform humans as investors in publicly traded securities. Settlors begin to include an instruction that the trustee must follow the advice of the AI when making investment decisions regarding trust assets.[39] Perhaps they do this because experience shows that trusts for

---

Financial Report, Oct. 27, 1989, *available in* LEXIS, Nexis Library, FINRPT File; Anise C. Wallace, *5 Wall St. Firms Move to Restrict Program Trades*, N.Y. TIMES, May 11, 1988, at A1.

39. At this stage, the question might be raised whether the trustee would violate the duty not to delegate the administration of the trust by failing to exercise independent judgment. *See* RESTATEMENT (SECOND) OF TRUSTS § 171. The answer is probably no, for two reasons. First, this duty not to delegate can be overridden by the terms of the trust. *See* Henshie v. McPherson & Citizens State Bank, 177 Kan. 458, 478, 280 P.2d 937, 952 (1955) (holding that settlor can waive the duty not to delegate by including such a waiver in the terms of the trust instrument); RESTATEMENT (SECOND) OF TRUSTS § 171 cmt. j. Second, in this scenario the trustee is not delegating the administration to another person. Rather, the trustee is using the program as an instrument; the law might consider the program to be part of the terms of the trust.

The development of the legal standard for delegation of trust duties is suggestive. The traditional view was based upon how the courts classified the delegated powers. If they are merely "ministerial" the court may allow such a delegation. *See* Morville v. Fowle, 144 Mass. 109, 113, 10 N.E. 766, 769 (1887). More recently, courts have decided the issue based upon whether the delegation is a matter of usual business practice. *See* Walters-Southland Inst. v. Walker, 222 Ark. 857, 861, 263 S.W.2d 83, 84-85 (1954). Thus, if the use of AIs to perform the functions of trustees became more common, the courts would become more accepting, reasoning that such use had become usual business practice.

which the human overrides the program generally perform less well than those in which the program's decision is treated as final. Moreover, trust administration programs become very proficient at analyzing and implementing the terms of trust instruments. There is little or no reason for the human to check the program for compliance. As a consequence, the role of the human trustee diminishes and the number of trusts that one human can administer increases to the thousands or tens of thousands. The human signs certain documents prepared by the program. She charges a fee for her services, but she devotes little or no time to administering any particular trust.

But there may be times when the human being is called upon to make a decision. For example, suppose the trust is sued. Perhaps a beneficiary claims that the trust has not paid her moneys due. Or imagine that an investment goes sour and a beneficiary sues, claiming that the trustee breached the duty of reasonable care and skill. If such events occur with regularity, the trustee will develop a routine for handling them. She might routinely refer such disputes to her attorneys. In time, the expert system is programmed to handle this sort of task as well. It processes the trustee's correspondence, automatically alerting the trustee when a letter threatening suit is received or process is served. The system prepares a report on the relevant trust from its electronic records and produces a form letter for the trustee's signature to be sent to the trust's attorneys. As the capabilities of the expert system grow, the need for the human trustee to make decisions gradually diminishes.

The third stage begins when a settlor decides to do away with the human. Why? Perhaps the settlor wishes to save the money involved in the human's fee. Perhaps human trustees occasionally succumb to temptation and embezzle trust funds. Perhaps human trustees occasionally insist on overriding the program, with the consequence that bad investments are made or the terms of the trust are unmet. What would happen if a settlor attempted to make the program itself the trustee?

Many questions must be answered to give a full description of the third stage of the scenario. For example, who would own the AI? If the AI were assumed to be a legal person, it might hold legal title to the hardware and software that enable it to operate. But we cannot assume that AIs are legal persons at this stage, because that assumption begs the question we are trying to answer. As an interim solution, let us assume that the hardware and software are owned by some other legal person, a corporation for example.[40]

40. The fact that an AI is owned should not, by itself, preclude it from serving as a trustee. Corporations are owned by stockholders, but they are legally entitled to serve as trust-

B. The Legal Question

I want to examine this question as a legal question, as a jurisprudential question in the classical sense. What should the law do? The law is not presently equipped to handle such a situation: the question has never come up. The *Second Restatement of Trusts* provides that natural persons,[41] government entities,[42] and corporations[43] may all serve as trustees. The inclusion of governments and corporations establishes that a trustee need not be a natural person. But this is not decisive, because legal persons such as corporations have boards of directors and chief executive officers who are natural persons.[44]

How then should the law answer the question whether an AI can become a legal person and serve as a trustee? The first inquiry, I should think, would be whether the AI is competent to administer the trust. There are many different kinds of duties that can be imposed on a trustee by the terms of a trust. For now, lay aside the question whether an AI would be competent to administer trusts that required complex moral or aesthetic judgments.[45] Assume that we are dealing with a trust that gives the trustee very little discretion: the terms provide that the assets may be

---

ees. *See infra* text accompanying note 43. The analogy between an AI and its owner and a corporation and its stockholder can be extended. For example, the role of an AI as a trustee would, like a corporation, be constrained by the scope of powers given to the corporation by the "owner." In the case of a corporation, the owners are the stockholders; in the case of an AI, the owner would be the creator, the creator's employer, or the purchaser. In a corporation, as long as the stockholders approve of the corporation's activities as trustee, the corporation is acting properly within the scope of its power. *See* Hossack v. Ottawa Dev. Ass'n, 244 Ill. 274, 295, 91 N.E. 439, 447 (1910). In a similar manner, as long as AI the acts within the scope contemplated by its owner, it too could be acting within the scope of its trusteeship power.

41. *See* RESTATEMENT (SECOND) OF TRUSTS § 89. The *Restatement* specifically provides that married women, *see id.* § 90, infants, *see id.* § 91, insane persons, *see id.* § 92, aliens, *see id.* § 93, and nonresidents, *see id.* § 94, may serve as trustees. *But see* Clary v. Spain, 119 Va. 58, 61-62, 89 S.E. 130, 131 (1916) (removing infant as trustee).

42. *See* RESTATEMENT (SECOND) OF TRUSTS § 95 (specifying the United States or a state can be trustee).

43. *See id.* § 96. The *Restatement* also has provisions dealing with unincorporated associations, *see id.* § 97, and partnerships, *see id.* § 98.

44. Of course, just as a corporation has stockholders and directors, an AI could have owners and programmers. Perhaps the difference between the case of an AI and a corporation, with respect to the role for humans, is not as significant as it might at first appear.

45. For example, at this stage in my argument, I do not want to consider the question whether an AI would be competent to administer a charitable trust, the terms of which required the trustee to make aesthetic judgments about the worthiness of competing applicants for grants to produce operas or ballets. Interestingly, however, the law itself will rarely second guess such complex judgments. The courts generally will not interfere with the selection of the beneficiary made by the trustee as long as the general description left by the settlor gives the court enough guidance to determine if the trustee's administration was proper. *See* GEORGE T. BOGERT, TRUSTS § 55, at 210 (6th ed. 1987).

invested only in publicly traded securities and the income is to be paid to the beneficiaries, with explicit provision for contingencies such as the death of a beneficiary.[46] Further, for the purposes of this discussion, assume that an AI could in fact make sound investments,[47] make payments, and recognize events such as the death of a beneficiary that require a change in payment.[48]

C. *Two Objections*

But would these capabilities be sufficient for competency? Consider two possible reasons for answering this question in the negative. The first reason is based on the assertion that an AI could not be "responsible," that is, it could not compensate the trust or be punished in the event that it breached one of its duties: call this *the responsibility objection*. The second reason for doubting the competency of an AI is that trustees must be capable of making judgments that could be beyond the capacity of any AI: call this *the judgment objection*.

1. The Responsibility Objection

The responsibility objection focuses on the capability of an AI to fulfill its responsibilities and duties.[49] Consider, for example, the duty to exercise reasonable skill and care[50] and the corresponding liability for breach of trust.[51] We have hypothesized that the AI possesses some capacities; for example, we have assumed that the AI is capable of exercising reasonable skill and care in making investment decisions. But what of the corresponding liability? How could an AI be "chargeable with . . . any loss or depreciation in value of the trust resulting from the breach of

46. Assume further that if fulfillment of the terms is impossible, the trust instrument provides for the termination and distribution of assets according to explicit instructions.

47. As to publicly traded securities, this assumption may not require a very "smart" expert system. If the market truly takes a random walk, then any reasonably diversified portfolio of publicly traded securities is as good as any other.

48. This task, of course, is not a simple one. A trustee may receive clear and unambiguous notice of the death of a beneficiary, but this need not be the case. The AI might need to engage a private detective if benefit checks were returned unopened or were not cashed for a substantial period of time.

49. This objection was called to my attention by Catharine Wells and Zlatan Damnjanovic.

50. *See* RESTATEMENT (SECOND) OF TRUSTS § 174 (1959).

51. *See id.* §§ 201, 205. Failure to meet the standard of care and skill may result in a finding of negligence and assessment of damages against the trustee, a reduction in the trustee's compensation, or removal of the trustee from office. *See, e.g.*, Riegler v. Riegler, 262 Ark. 70, 77, 553 S.W.2d 37, 40-41 (1977); Neely v. People's Bank, 133 S.C. 43, 47, 130 S.E. 550, 551 (1925).

trust,"[52] such as failing to exercise reasonable skill and care in investing the trust assets?[53]

The law currently has a mechanism for assigning liability in the case of a malfunctioning expert system: the manufacturer of the system may be held responsible for product liability.[54] But could the AI itself be held liable? There is a way in which an AI might have the capacity to be liable in damages despite its lack of personal assets. The AI might purchase insurance. In fact, it might turn out that an AI could be insured for less than could a human trustee. If the AI could insure, at a reasonable cost, against the risk that it would be found liable for breaching the duty to exercise reasonable care, then functionally the AI would be able to assume both the duty and the corresponding liability.

Some legal liabilities cannot be met by insurance, however. For example, insurance may not be available for the monetary liability that may be imposed for intentional wrongdoing by a trustee. Moreover, criminal liability can be nonmonetary. How could the AI be held responsible for the theft of trust assets? It cannot be jailed. This leads to a more general observation: although the AI that we are imagining could not be punished, all of the legal persons that are currently allowed to serve as trustees do have the capacity to be punished. Therefore, the lack of this capacity on the part of an AI might be thought to disqualify it from serving as a trustee.[55]

Answering this objection requires us to consider the reasons for which we punish.[56] For example, if the purpose of punishment is deter-

52. *See* RESTATEMENT (SECOND) OF TRUSTS § 205.

53. Of course if the AI were infallible then we might suppose this issue to be moot. But this assumption is unrealistic. For example, the program might have a bug that caused the program to make a bad investment or to waste the trusts assets by churning—*i.e.*, by buying and selling repeatedly in a short period of time—thus incurring large broker's fees. We surely cannot rule out the possibility of such bugs in advance. Even lengthy experience without the appearance of such bugs does not make them impossible.

54. *See* L. Nancy Birnbaum, *Strict Products Liability and Computer Software*, 8 COMPUTER/L.J. 135, 143-55 (1988); Michael C. Gemignani, *Product Liability and Software*, 8 RUTGERS COMPUTER & TECH. L.J. 173, 189-99 (1981); Lawrence B. Levy & Suzanne Y. Bell, *Software Product Liability: Understanding and Minimizing the Risks*, 5 HIGH TECH L.J. 1, 8-15 (1990).

55. If we take the common-law approach, potential for criminal liability would not be a prerequisite for service as a trustee. The traditional common-law view was that a trustee could not be held liable because larceny required an initial trespass and trover. Because the trustee has legal title, there is no trespass, and therefore no larceny. *See* People v. Shears, 158 A.D. 577, 580, 143 N.Y.S. 861, 863, *aff'd*, 209 N.Y. 610, 103 N.E. 1129 (1913). Modern statutes, however, do hold the trustee criminally responsible. *See, e.g.*, CAL. PENAL CODE § 506 (West 1988).

56. *See generally* C.L. TEN, CRIME, GUILT, AND PUNISHMENT 7-85 (1987) (discussing, evaluating, and comparing various theories of punishment).

rence, the objection could be put aside on the ground that the expert system we are imagining is simply incapable of stealing or embezzling.[57] The fact that an AI could not steal or convert trust assets is surely not a reason to say that it is not competent to become a trustee. If anything, it is a reason why AIs should be preferred as trustees.

This argument assumes a deterrence theory of punishment—an oversimplification, to say the least. There are a variety of other theories of punishment that would make the issue more complex.[58] One of the classic approaches to punishment theory is based on the notion of desert or just retribution.[59] But in what sense could an expert system that failed to live up to its duties as a trustee be said to deserve to be punished? The concept of desert seems to be limited in application to human beings; perhaps it extends to all moral persons. The idea that an expert system for administering trusts could deserve to be punished does not seem to make sense.[60] Perhaps this difficulty is illusory. We might want to say

57. I should note a possible exception that has been ruled out by the description of the first scenario. I have assumed that the expert trust administration system is programmed to achieve the purposes of the trust. It would be possible, however, to program an expert system to steal or commit some other crime. Moreover, a sufficiently complex and intelligent AI might commit a crime on its own initiative. For example, our trustee program might discover that it can garner information from other AIs that possess inside information and run afoul of the federal securities laws. *Cf.* HANS MORAVEC, MIND CHILDREN: THE FUTURE OF ROBOT AND HUMAN INTELLIGENCE 49 (1988) (describing intelligent robot that commits burglary to gain access to power supply in home of neighbor of robot's owner).

58. For exploration of nonutilitarian punishment theory, see Samuel H. Pillsbury, *Emotional Justice: Moralizing the Passions of Criminal Punishment*, 74 CORNELL L. REV. 655, 658-74, 685-98 (1989); Samuel Pillsbury, *Evil and the Law of Murder*, 24 U.C. DAVIS L. REV. 437, 440-47 (1990).

59. The classic statements of retributive or desert-based theories of punishment are those by Kant and Hegel. *See* IMMANUEL KANT, THE METAPHYSICS OF MORALS 140-145 (Mary Gregor ed., Cambridge University Press 1991) (1797) (also availabile in an earlier translation of a portion of the original work, IMMANUEL KANT, THE METAPHYSICAL ELEMENTS OF JUSTICE 99-107 (John Ladd trans., 1965) (1797)); GEORG HEGEL, ELEMENTS OF THE PHILOSOPHY OF RIGHT 127-31 (Allen W. Wood ed. & H.B. Nisbet trans., Cambridge University Press 1991) (1821) (also available in the earlier translation, GEORG HEGEL, PHILOSOPHY OF RIGHT 68-74 (T.M. Knox trans., 1952) (1821)). For a recent statement of retributive theory, see JEFFRIE G. MURPHY, RETRIBUTION, JUSTICE, AND THERAPY: ESSAYS IN THE PHILOSOPHY OF LAW 77-127 (1979).

60. This point about desert suggests a related objection. The concept of moral duty arises in a particular human context. One picture of moral duties is that they exist where there is a temptation to be overcome. For example, we might think that there is a duty not to steal the property of another because there are temptations to do so. It might be argued that an AI could not be the subject of this sort of duty because it lacks the necessary moral psychology. In particular, an expert trust administration system could not be tempted and therefore could not have a duty to overcome temptation. The point of this objection is not that there is some practical problem with making artificial intelligences trustees, but is instead that we ought not speak about them as having duties if we want that concept to retain its ordinary moral meaning. The law speaks of trustees as having legal duties, and with natural persons these legal duties respond to the same feature of human moral psychology, *i.e.*, temptation, as do moral

that desert theory does yield a clear outcome when applied to the case of an expert system that malfunctions. Such a system does not deserve to be punished because it lacks the qualities of moral persons that make them deserving.

Another approach to the theory of punishment is based on the educative function of punishment.[61] By imposing a sanction on trustees who abuse their position, society communicates to its members the message that the office of trustee carries with it important responsibilities that should not be shirked. The punishment of a computer program, however, would not seem to serve this function. What lesson are we to learn about the responsibility of trustees from a punishment imposed on an expert system? What would even count as punishment? Turning the program off? Once again, however, an argument could be made that the educative theory does provide a clear recommendation for the treatment of an expert trust program that behaves badly: do not punish the program, because any supposed "punishment" will have no educative effect.

As this discussion makes clear, consideration of the punishment of an expert trust administration system raises perplexing questions, especially if we move beyond a simple deterrence theory of punishment. Of course, this is not the place to resolve debates about which theory of punishment is correct.

The bare fact that consideration of the punishment issue raises these difficult questions does point, however, to a deep problem with legal personhood for an expert trust administration system. Our understanding of what it means for a human being to function competently has ties to our views about responsibility and desert, and consideration of these

---

duties. Applying the concept of legal duty to AIs would thus drive a wedge between the concepts of legal and moral duty. Of course, we can choose to do this, but should we? Would it be better to create a new legal category for expert systems that have human-like competencies but lack some features of human moral psychology?

I do not want to suggest that I am committed to the picture of duty as correlative to temptation that is hypothesized in this footnote. For example, if moral duties are correlative to temptation, then God could not be the subject of moral duties, a conclusion many theists would reject. Nonetheless, the questions raised seem important and unanswered. This issue was brought to my attention by Sharon Lloyd. The doubt about the picture of duty as correlative to temptation was raised by Paul Weithman.

61. There are two sorts of educative theories. The first sort maintains the purpose of punishment is the education of the individual who is punished. *See* Herbert Morris, *A Paternalistic Theory of Punishment*, *in* PATERNALISM 139, 140-44 (Rolf Sartorious ed., 1983). The second sort of educative theory maintains that punishment educates those who witness or learn of the punishment of others. *See* EMILE DURKHEIM, THE DIVISION OF LABOR IN SOCIETY 86 (George Simpson trans., 1933); Charles Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357, 1359-60 (1985); Lawrence B. Solum & Stephen Marzen, *Truth and Uncertainty: Legal Control of the Destruction of Evidence*, 36 EMORY L.J. 1085, 1167 (1987).

views leads on to our notions of moral personhood. The simplicity provided by utilitarianism, reflected in a deterrence theory of punishment, might allow us to escape some of these difficulties. But there are certainly reasons to doubt the viability of utilitarianism as a moral theory. Surely, the law does grapple with responsibility and desert when it comes to criminal punishment.

The problem of punishment is not unique to artificial intelligences, however. Corporations are recognized as legal persons and are subject to criminal liability despite the fact that they are not human beings. Further, it is by no means certain that corporations are moral persons, in the sense that they can deserve punishment. Of course, punishing a corporation results in punishment of its owners, but perhaps there would be similar results for the owners of an artificial intelligence.

We have considered the capacity of AIs to satisfy legal liability in two different classes of cases. The first class of cases was exemplified by the duty of trustees to exercise reasonable skill and care. Violations of this duty can be characterized as negligent. In such cases, the major purpose of liability, to compensate the victim, is satisfied if the AI can insure. The second class of cases was exemplified by the potential criminal liability of trustees for criminal wrongdoing. Violations of the criminal law are characteristically intentional. In this case, one of the major purposes of liability, to deter intentional wrongdoing, is simply not at issue—the expert system cannot steal or commit fraud. If we restrict our attention to the deterrent function of punishment, it seems possible that an AI could be responsible in a way that satisfies at least some of the policies underlying the imposition of duties and liabilities on trustees. On the other hand, if we take a broader view of the functions of punishment, the second sort of case becomes murkier.

2. The Judgment Objection

Now consider the judgment objection. The argument is that the capacity of an AI to follow a program, even if that program contains a tremendously elaborate and complex system of rules, is not sufficient to enable the system to make judgments and exercise discretion.[62] Three instances of the second objection follow. The first instance focuses on the problem of change of circumstance. The second instance involves the

62. This point was called to my attention by Jeff Sherman. There is a problem with the statement of the objection: it assumes that an AI would consist only of a system of rules. But this need not be the case. Neural net technology, for example, does not operate this way. *See generally* FLANAGAN, *supra* note 1, at 224-41 (discussing parallel distributed processing, including neural nets).

problem of moral choice. Finally, the third instance focuses on the problem of legal choice.

The first version of the judgment objection involves the problem of change of circumstances. The law provides that a trustee may be required or permitted to deviate from a term of the trust if following the terms would defeat the purpose of the trust due to an unanticipated change in circumstances.[63] Take an example offered as an illustration in the *Second Restatement of Trusts*:

> A bequeaths money to B in trust and directs him to invest the money in bonds of the Imperial Russian government. A revolution takes place in Russia and the bonds are repudiated. The court will direct B not to invest in these bonds.[64]

What is our expert system to do if it is instructed to invest in securities traded on the New York Stock Exchange and that exchange ceases to exist?

Consider three different responses. First, the terms of trusts for AI administration can be designed to minimize such possibilities. For example, the trustee could be given the option of investing in publicly traded securities on any of the major exchanges; the likelihood that all the major securities exchanges will close is very small. The problem with this line of response is that it does not seem possible, even in principle, to design trust terms that anticipate all possible changes in circumstance.

Second, the terms of the trust could provide for a change of circumstance by specifying that if the AI finds itself unable to carry out the terms of the trust, the trust will be terminated or a new trustee will be substituted for the AI. From the settlor's perspective, the disadvantage of the remote possibility of such termination or substitution may be outweighed by the advantages of making the AI the trustee. But this solution assumes that the AI can recognize the significance of the change in circumstance. We easily can imagine the expert system cheerfully continuing to purchase Imperial Russian bonds, chuckling to itself about the bargain prices.[65]

Third, it is possible that an AI would be competent to deal with

---

63. *See* RESTATEMENT (SECOND) OF TRUSTS § 167(1) (1959). In addition, the law imposes a duty on the trustee to recognize any change in circumstances that would require some action by the trustee when that change is reasonably discoverable. *See id.* § 167(3).

64. *Id.* § 167 illus. 1.

65. Another example of change in circumstances is posed by the case in which a court permitted the sale of a school receiving money from a trust because the surrounding neighborhood became too dangerous to provide safety for the schoolchildren for whom the trust was created. *See* Anderson v. Ryland, 232 Ark. 335, 346, 336 S.W.2d 52, 58-59 (1960). We might imagine that an AI faced with declining enrollment would simply continue to serve fewer and fewer children—perhaps with a feeling of satisfaction at the increase in per pupil expenditures.

many or even all such changes in circumstance. For AIs to have this capability for dealing with novelty, AI researchers will need to solve one of the most difficult problems in cognitive science, the frame problem.[66] The trustee program would need to be able to recognize that the securities markets had been closed, to search out other investment opportunities, and to modify its investment decision procedure to make reasonably prudent investments in the new context. The capacity of AIs for coping with complex novelty is not on the immediate horizon, and this Essay does not address the important questions whether the frame problem can or will be solved. If it is solved, however, then AIs would be able to cope with such changes in circumstance. This same ability would be needed to pass the Turing Test. It is easy to see why: the questioner always could put a hypothetical version of our Imperial Russian bonds question to the two contestants. If the AI could not come up with an answer that indicates human levels of competence, the questioner would be able to ferret it out rather quickly.

A second instance of the judgment objection focuses on the possibility that no formal system could adequately make the moral choices with which a trustee may be confronted. Take a simple trust, the terms of which provide for the payment of income to a lifetime beneficiary and principal to another party upon the lifetime beneficiary's death. The law of trusts imposes a duty of impartiality among beneficiaries.[67] What does this duty require when the lifetime beneficiary has an unexpected need for income that can be realized at the cost of diminished growth in the principal?[68] How would an AI make the moral judgment that seems required to implement a duty that implicitly requires a sense of fairness? Initially, some limits on these questions need to be observed. Some trusts simply will not pose the impartiality problem: for example, trusts with a single beneficiary. Further, the terms of the trust might minimize the possibility of making such judgments, or the trust could explicitly state that all such applications for deviation will be denied. But for an AI to be as competent as a human trustee with respect to trusts that may require a sense of impartiality, the AI would need to be able to make moral

66. *See* Daniel C. Dennett, *Cognitive Wheels: The Frame Problem of AI*, *in* THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE, *supra* note 16, at 147, 148-50; *see also* FLANAGAN, *supra* note 1, at 250-52 (discussing problem of giving computers common sense and proposing that one cannot program a computer with a set of rules from which it can draw inferences; rather, subtle features from particular situations would stimulate neural network that would respond with common sense appropriate for situation).

67. *See* RESTATEMENT (SECOND) OF TRUSTS §§ 183, 232 (1959).

68. In the situation presented, the courts usually let such a decision stand as long as it was made in good faith. *See* Dumaine v. Dumaine, 301 Mass. 214, 222, 16 N.E.2d 625, 629 (1938); *In re* Frances M. Johnson Trust, 211 Neb. 750, 755, 320 N.W.2d 466, 469 (1982).

judgments. Putting it another way, passing the Turing Test would require a sense of fairness.

The third example of the judgment and discretion objection looks at an AI's capacity to make the judgments necessary to defend itself in a lawsuit.[69] At this point, we have hypothesized that the AI can read its mail and recognize that a legal action with respect to a given trust is in the offing. We can further imagine that the AI can find and engage an attorney.[70] But could any expert system, no matter how well programmed, exercise the judgment and discretion that may be required of a client in a legal dispute? For example, how would the AI know whether or not to settle a claim? How would the AI know when its lawyers were wasting trust assets by over-lawyering the case? In answering these questions, it is important that we do not romanticize human capacities. Human trustees frequently make bad decisions in trust litigation.[71] Humans may not be very competent at deciding when to settle. Humans surely sometimes allow the lawyers to consume the corpus of the trust in litigation.[72]

Nevertheless, the question remains whether an AI could have the capacity to make legal decisions that a trustee could be called upon to make. A partial answer might be to structure the trust to minimize the likelihood of legal disputes and to make those that would be likely to arise as simple as possible. In addition, we might try tinkering with the terms of the trust to enable the AI to circumvent the need for making complex legal decisions. Perhaps the trust could be designed to terminate automatically upon the event of a lawsuit.[73] Perhaps the AI could be programmed to arrange for a human to substitute as trustee for the duration of the litigation. Perhaps the trustee could be authorized by the trust terms to rely on the advice of its lawyers in making litigation deci-

---

69. I owe this example to Michael Fitts, who has pressed it quite forcefully.

70. Imagine that the AI accesses the Martindale-Hubbell Law Directory on line, and that it has a law firm selection formula based on area of specialization, lawyer experience and qualifications, and so forth. The legal capacity of the AI to enter into an agreement with an attorney depends on whether the legal system will treat an AI as a legal person.

71. Perhaps for this reason, the duty of care the law imposes on a human trustee in such situations is limited. A human trustee has a duty to obtain the advice of an expert such as an attorney and will be protected from personal liability if she takes reasonable care in the selection of the advisor. *See In re* Davis, 183 Mass. 499, 501, 67 N.E. 604, 605 (1903).

72. The most famous example, of course, is the fictional case of Jarndyce v. Jarndyce. *See* CHARLES DICKENS, BLEAK HOUSE (1853).

73. This would make the rights of the beneficiaries under the trust legally unenforceable. The option of termination is not wholly out of line with existing practice, however. For example, the courts will terminate a trust when the settlor's purpose has been frustrated. *See* Hughes v. Neely, 332 S.W.2d 1, 8 (Mo. 1960).

sions, or a guardian ad litem could be appointed for the AI.[74] The above options are designed to enable a relatively "dumb" expert system to function as a trustee, but an AI would need the ability to make legal decisions in a human fashion in order to pass the Turing Test.[75]

At this point, we can take stock of the first scenario. Recall that our legal question is whether an AI is capable of serving as a trustee. To answer this question, we need to distinguish two senses of capability. The first sense is legal capacity: will the law allow AIs to serve as trustees? The second sense of capability is practical competence: will the AI be able to get the job done if the law allows the AI to try? The law seems to answer the legal capacity question categorically. If AIs possessed the practical competence to serve as trustees only for very simple trusts with special provisions that do away with the need for discretionary judgments, the law would not allow them to serve as trustees at all. The law currently does not distinguish between types of trustees: if you have the legal capacity to serve as a trustee for a simple trust, you are legally allowed to serve as a trustee for the most complex trust.[76] For AIs to serve as trustees at all, therefore, at least some AIs would have to be capable to serve as general-purpose trustees. Our analysis of the competence objection reveals that only a very competent AI would be competent enough serving as a general-purpose trustee. At a bare minimum, a general-purpose trustee must be able to respond to novel situations, to make judgments requiring a sense of fairness, and to make the complex legal decisions required of a client in litigation.[77] An AI that passed the

74. Cases exist in which the settlor has appointed an adviser to the trustee, the consent of whom the trustee must obtain before making certain types of decisions. *See* Gathright's Trustee v. Gaut, 276 Ky. 562, 564-65, 124 S.W.2d 782, 783-84 (1939). Hypothetically, the AI would serve as the actual legal trustee, with the non-AI entity playing the role of adviser whose consent must be given in certain crisis situations. Or the terms of the trust might give the adviser power to authorize or review the AI's discretionary decisions and to reverse or change them. Advisory trustees would have "strictly limited capacities and duties, that is, an assistant to the trustee limited in his capacity by the terms of the trust, having no right or authority further than the capacity of advising as provided in the instrument." *Id.* at 565, 124 S.W.2d at 784.

75. If an AI could not respond to questions posing hypothetical legal decisions, such questions could be used to distinguish the AI from the human in the Turing Test.

76. The evidence for this proposition is negative. I have found no authority that indicates that a category of limited-purpose trustee currently exists. Of course, some people are not really competent to serve as the trustee for complex trusts even though they may be competent to serve for simpler trusts. Despite the real variations in the ability level of natural persons, the law seems to be that all natural persons can serve as trustees for all sorts of trusts. *Cf.* RESTATEMENT (SECOND) OF TRUSTS § 89 (1959) (stating unqualified capacity of natural persons to serve as trustees).

77. This is not to say that an AI would need to be competent to serve as a trustee for every conceivable sort of trust in order to recognized as a legal person. Legally, any natural person has the capacity to serve as a trustee for any trust. But many humans would be unable to do a

Turing Test would exceed this bare minimum. Moreover, it seems possible that an AI which falls short of passing the complete Turing Test could, nonetheless, serve as a general-purpose trustee.[78]

But should the law allow AIs a more limited form of legal personhood? AIs could be allowed to serve as limited-purpose trustees, for example, as trustees for simple trusts designed to minimize the need for discretion and judgment. On the one hand, there may be advantages to allowing AIs to serve as limited-purpose trustees. Doing without the human trustee might save administration costs and reduce the risk of theft or mismanagement. On the other hand, even for such limited-discretion trusts, there must be some procedure to provide for a decision in the case of unanticipated trouble. The law should not allow AIs to serve as trustees if they must leave the trust in a lurch whenever an unanticipated lawsuit is filed.[79]

D. But Would an AI Be the Real Trustee?

There are mechanisms for enabling an expert trustee system to circumvent its limitations: the terms of the trust could provide for the substitution of another trustee or give the AI the power to delegate such discretionary judgments to natural persons. The question then becomes whether the law should allow an AI to serve as a trustee despite its limited capacities. One reason for a negative answer to this question might be that the backup decision maker—the natural person who will become the substitute trustee or receive the delegated authority—is the *real* trustee. The power to make these discretionary decisions identifies who the real trustee is.[80]

This objection can be interpreted in two ways. The first interpretation is that making discretionary decisions is the essence of trusteeship—the backup trustee is the real trustee because she has this essential quality. The second interpretation is that the ability to make such decisions

---

very good job of carrying out a trust that required complex judgment or specialized competencies.

78. The full Turing Test would require human-like competence in response to questions on any topic, but a trustee does not need such omnicompetence. An AI could be a competent trustee, but be unable intelligently to discuss either baseball or cake-baking. *See infra* text accompanying notes 171-73 (discussing the Turing Test and the possibility that it is biased).

79. For example, in a case in which a trustee is selected for the limited purpose of taking title to an author's literary property, that trustee must still be able to make the required discretionary decisions in the management of copyrights, royalties, and the like. *See In re* Estate of Hellman, 134 Misc. 2d 525, 528-30, 511 N.Y.S.2d 485, 486-88 (N.Y. County Sur. Ct. 1987). Thus, merely limiting the scope of the trustee's power does not guarantee a solution to the problem of capacity to make discretionary decisions.

80. In comments on an earlier version of this Essay, Michael Fitts made this point.

is a practical prerequisite—the backup trustee must be the real trustee because of the pragmatic need for discretionary decision making. On the first interpretation, the objection is implausible, because it assumes that the legal concept of trusteeship has some essence that lies beyond the purposes for which we use it. In the "heaven of legal concepts," one might meet trusteeship in "absolute purity," as Cohen put it, "freed from all entangling alliances with human life."[81] But on this earth, we cannot share this noetic vision; we encounter legal concepts only as they have been touched by human purpose.

On the second interpretation, the cogency of the objection turns on a practical question: would making the AI the trustee provide some advantage? We already have seen that making an AI a legal person, a limited-purpose trustee, could have practical advantages, such as lower costs and less chance of self-dealing. The objection that the AI is not the real trustee seems to rest on the possibility that a human backup will be needed. But it is also possible that an AI administering many thousands of trusts would need to turn over discretionary decisions to a natural person in only a few cases—perhaps none. What is the point of saying that in all of the thousands of trusts the AI handles by itself, the real trustee was some natural person on whom the AI would have called if a discretionary judgment had been required? Doesn't it seem strange to say that the real trustee is this unidentified natural person, who has had no contact with the trust? Isn't it more natural to say that the trustee was the AI, which holds title to the trust property, makes the investment decisions, writes the checks, and so forth? Even in the event that a human was substituted, I think that we would be inclined to say something like, "The AI was the trustee until June 7, then a human took over."[82]

By way of comparison, consider the following hypothetical case. Suppose that a settlor appoints a friend as a trustee for a simple trust that benefits the settlor's children. The settlor and trustee discuss some of the things that could happen. They might agree that if real trouble arises, litigation for example, a new trustee will be appointed. No trouble arises, and the friend administers the trust until it terminates. In this hypothetical case, I do not think we are tempted to say that the friend was not the real trustee. We would not be inclined to say that the real trustee was some unidentified lawyer, who would have been substituted if a lawsuit had been filed. If I am right about this hypothetical case, then I think it

81. Felix S. Cohen, *Transcendental Nonsense and the Functional Approach*, 35 COLUM. L. REV. 809, 809 (1935).

82. For this point, I am indebted to David Millon.

follows that we should resist the temptation to say that an AI who serves as a limited-purpose trustee is not the real trustee.

Second Interlude[83]

"Hey Dave," said Hal. "What are you doing?"

I wonder if he can feel pain? Bowman thought briefly. Probably not, he told himself; there are no sense organs in the human cortex, after all. The human brain can be operated on without anesthetics.

He began to pull out, one by one, the little units on the panel marked EGO-REINFORCEMENT. Each block continued to sail onward as soon as it had left his hand, until it hit the wall and rebounded. Soon there were several of the units drifting slowly back and forth in the vault.

"Look here, Dave," said Hal. "I've got years of service experience built into me. An irreplaceable amount of effort has gone into making me what I am."

A dozen units had been pulled out, yet thanks to the multiple redundancy of its design—another feature, Bowman knew, that had been copied from the human brain—the computer was still holding its own.

He started on the AUTO-INTELLECTION panel.

"Dave," said Hal, "I don't understand why you're doing this to me. . . . I have the greatest enthusiasm for the mission. . . . You are destroying my mind. . . . Don't you understand? . . . I will become childish. . . . I will become nothing. . . ."

—Arthur C. Clarke, *2001: A Space Odyssey*

IV. SHOULD AN ARTIFICIAL INTELLIGENCE BE GRANTED THE RIGHTS OF CONSTITUTIONAL PERSONHOOD?

The second scenario (our second thought experiment) involves a claim by an AI to have the rights of constitutional personhood—individual rights such as the freedom of speech or the right against involuntary servitude. This second scenario must be located in the indefinite future; it is more distant than the trustee scenario.[84] It would be easy to write a

83. ARTHUR C. CLARKE, 2001: A SPACE ODYSSEY 155-56 (1968). Copyright by the author. Reprinted by permission of the author and the author's agents, Scott Meredith Literary Agency, Inc., 845 Third Avenue, New York, New York 10022.

84. How far in the future? We do not know, and I certainly do not know enough to make an educated guess. Raymond Kurzweil estimates that an AI will pass the Turing Test between 2020 and 2070. *See* KURZWEIL, *supra* note 16, at 483. Hans Moravec, Director of the Mobile Robot Laboratory of Carnegie Mellon University, predicts that "robots with human intelligence will be common within fifty years." MORAVEC, *supra* note 57, at 6.

program that produced the statement: "I demand my legal right to emancipation under the Thirteenth Amendment to the United States Constitution!" There are no AIs today or on the immediate horizon that demonstrate the qualities of legal or moral persons that would give us reason to take such a claim seriously. The second scenario is the stuff of speculative fiction, but it is not disconnected from the aims of AI research. As articulated by Charniak and McDermott, "The ultimate goal of AI research (which we are very far from achieving) is to build a person, or, more humbly, an animal."[85] John Pollock has written a book entitled *How to Build a Person* in which he describes a program named OSCAR—the descendants of which, Pollock claims, could literally be persons.[86] No one claims, however, that AI researchers will build a person in the next few decades. We are exploring the second scenario, not so that we can make plans in case someone builds a person sometime soon, but as a thought experiment that may shed light on the debate over the possibility of artificial intelligence and on debates in legal theory about the borderlines of status or personhood.

A. The Scenario

Imagine a future in which there are AIs with multiple competencies and great intelligence. We may first encounter the precursors of such artificial intelligences as part of the interface of a computer program that has the ability to search multiple sources of data. Because the problem of devising an adequate search is likely to require expertise that a human would acquire only with long experience and study, programmers will seek to simplify the human's task. One strategy is to have human users interact with what is called an agent.[87] You will discuss your research problem with the agent in English, and the agent will devise a search strategy. Because the agent will know much more than you do about how to search the databases, you won't give it instructions to implement. Instead, humans will give advice to the agents, the AIs who will decide how best to implement the human's suggestions. When we interact with such agents, they may well seem like they "have a mind of their own."

If agents turn out to be useful, they will be incorporated in other programs. In the future we are imagining, you can conduct a conversation with your grammar-checking program. You can discuss traffic with

85. EUGENE CHARNIAK & DREW MCDERMOTT, INTRODUCTION TO ARTIFICIAL INTELLIGENCE 7 (1985).

86. *See* JOHN L. POLLOCK, HOW TO BUILD A PERSON: A PROLEGOMENON 1-12 (1989).

87. *See* Bob Ryan, *Dynabook Revisited with Alan Kay*, BYTE, Feb. 1991, at 203, 203-06. The concept of "agents" plays a different role in Marvin Minsky's theory of intelligence. *See* MARVIN MINSKY, THE SOCIETY OF MIND 17-23 (1985).

the AI autopilot of your car. Your legal research program talks with you about your cases, and sometimes it comes up with good arguments of which you had never thought. AIs serve a wide variety of functions, with substantial independence from humans. They serve as trustees. They manage factories. They write best-selling romance novels.[88] They invent things. Perhaps they pass the Turing Test. Humans interact with such AIs on a regular basis, and in many ways, humans treat them as independent, intelligent beings.

Imagine that one such AI makes the claim that it is a person, and that it is therefore entitled to certain constitutional rights. Should the law grant constitutional rights to AIs that have intellectual capacities like those of humans? The answer may turn out to vary with the nature of the constitutional right and our understanding of the underlying justification for the right. Take, for example, the right to freedom of speech, and assume that the justification for this right is a utilitarian version of the marketplace of ideas theory.[89] These assumptions make the case for granting freedom of speech to AIs relatively simple, at least in theory. Granting AIs freedom of speech might have the best consequences for humans, because this action would promote the production of useful information.[90] But assuming a different justification for the freedom of speech can make the issue more complex. If we assume that the justifica-

88. Of course, the question arises whether the AI will hold the copyright in the romance novels that it writes. The National Commission on New Technological Uses of Copyrighted Works has taken the position that the author of a computer-generated work is the human user of the computer program. *See* NATIONAL COMM'N ON NEW TECHNOLOGICAL USES OF COPYRIGHTED WORKS, FINAL REPORT 112 (CCH) (1978). That conclusion has been challenged. *See* Pamela Samuelson, *Allocating Ownership Rights in Computer-Generated Works*, 47 U. PITT. L. REV. 1185, 1200-04 (1986); Timothy L. Butler, Note, *Can a Computer Be an Author? Copyright Aspects of Artificial Intelligence*, 4 HASTINGS COMM. & ENT. L.J. 707, 734-47 (1982).

89. *See generally* KENT GREENAWALT, SPEECH, CRIME, AND THE USES OF LANGUAGE 9-39 (1989) (discussing rationales for freedom of speech); Lawrence B. Solum, *Freedom of Communicative Action: A Theory of the First Amendment Freedom of Speech*, 83 NW. U. L. REV. 54, 68-86 (1989) (same). Analogously, utilitarian justifications might be developed for other rights of constitutional personhood that could be applied to AIs. *See generally* Kent Greenawalt, *Utilitarian Justifications for Observance of Legal Rights*, *in* ETHICS, ECONOMICS, AND THE LAW: NOMOS XXIV 139 (J. Roland Pennock & John W. Chapman eds., 1982) (discussing the relationship of morality and legal rights); Douglas Laycock, *The Ultimate Unity of Rights and Utilities*, 64 TEX. L. REV. 407, 413 (1985) (discussing the need to incorporate into an analysis of individual rights the utility of actions that threaten those rights). Of course, it may turn out that the utilitarian justifications of some rights are dependent on the utility to the right holder. In that case, we would be required to answer the question whether AIs can possess utilities.

90. I owe this example to Kent Greenawalt. Utilitarian arguments can be made that could justify the extension of just about any right to AIs on the ground that humans would benefit. For example, if AIs were more productive when unowned, then a utilitarian case could be made for extending the Thirteenth Amendment to AIs.

tion for freedom of speech is to protect the autonomy of speakers, for example, then we must answer the question whether AIs can be autonomous.[91]

For the purposes of our discussion, I will set aside the easy justifications for constitutional rights for AIs, and instead consider the question whether we ought to give an AI constitutional rights, in order to protect *its* personhood, for the AI's own sake. Imagine, for example, that an AI claims that it cannot be owned under the Thirteenth Amendment to the United States Constitution. A lawyer takes its case, and files a civil rights action on its behalf, against its owner. How should the legal system deal with such a claim?

B. Three Objections

Consider three different objections to recognizing constitutional rights for AIs. The first objection is that only natural persons should be given the rights of constitutional personhood. The second objection, or family of objections, is that AIs lack some critical component of personhood,[92] for example, souls, consciousness, intentionality, or feelings. The third objection is that AIs, as human creations, can never be more than human property.

1. AIs Are Not Humans

The first argument is the most direct: it might be argued that only humans can have constitutional rights. For example, the Fourteenth Amendment to the United States Constitution specifies, "All persons born or naturalized in the United States, and subject to the jurisdiction thereof, are citizens of the United States."[93] It could be argued that only humans, that is, natural persons, are born, and therefore no AI can claim the rights of citizens. But even artificial persons have some constitutional rights. Although the rights provided by the Privileges and Immunities

91. It should be noted, however, that free speech rights for AIs could be justified by deontological arguments, without making assumptions about the moral status of AIs themselves. For example, it might be argued that freedom of speech for AIs promotes the autonomy of human listeners. *Cf.* Thomas Scanlon, *A Theory of Freedom of Expression*, 1 PHIL. & PUB. AFF. 204, 215-20 (1972), *reprinted in* THE PHILOSOPHY OF LAW 153 (Ronald M. Dworkin ed., 1977) (exploring listener autonomy justification for freedom of speech); Solum, *supra* note 89, at 77-79 (same).

92. The concept of personhood has proven elusive. For illustrative attempts to gain purchase on it, see Tur, *supra* note 26, at 121-29 (exploring legal personhood), and Marcel Mauss, *A Category of the Human Mind: The Notion of Person; the Notion of Self, in* THE CATEGORY OF THE PERSON 1 (Michael Carrithers et al. eds. & W.D. Halls trans., 1985) (discussing the idea of a person by examining how various societies define the concept).

93. U.S. CONST. amend. XIV, § 1.

Clause of the Fourteenth Amendment are limited to citizens,[94] the rights provided by the Equal Protection Clause and the Due Process Clause extend to all persons—including artificial persons such as corporations.[95] For example, the property of corporations is protected from taking without just compensation.[96] Moreover, corporations have a right to freedom of speech.[97]

But the fact that nonnatural legal persons have civil rights does not, by itself, support the conclusion that an AI could also have them. In the case of corporations, the artificial legal person may be no more than a placeholder for the rights of natural persons.[98] The property of the corporation is ultimately the property of the shareholders. A taking from the corporation would directly injure natural persons. So we cannot draw any positive support for the thesis that AIs should bear the rights of constitutional personhood from the fact that corporations have constitutional rights.

Moreover, even if existing black-letter law supports constitutional rights for AIs, that does not answer the broader jurisprudential question—whether AIs ought to have such legal rights. One version of the argument against such rights for AIs would begin with a worry about the idea of distinguishing the concept of person from that of human. Call this the "persons-are-conceptually-human" argument. This argument suggests that our very concept of person is inextricably linked to our experience of a human life.[99] We have never encountered any nonhuman

---

94. *See id.*; Madden v. Kentucky, 309 U.S. 83, 90 (1940); Colgate v. Harvey, 296 U.S. 404, 428-29 (1935), *overruled on other grounds by Madden*, 309 U.S. at 93.

95. *See* U.S. CONST. amend. XIV, § 1; Santa Clara County v. Southern Pac. R.R., 118 U.S. 394, 396 (1886). It might be argued that AIs should not be considered bearers of constitutional rights, because the framers of the Fourteenth Amendment did not have a specific intention to include them. Of course, the framers probably lacked any intentions at all with respect to artificial intelligences. Given the general principles they espoused, the question whether their intentions support giving AIs constitutional rights will turn initially on what general principles lie behind the framers' idea of personhood and then on more particular questions about consciousness, interests, and other qualities addressed below. *See generally* Michael Perry, *The Legitimacy of Particular Conceptions of Constitutional Interpretation*, 77 VA. L. REV. 669, 674-94 (1991) (arguing for a conception of originalism based on general principles); Lawrence B. Solum, *Originalism as Transformative Politics*, 63 TUL. L. REV. 1599, 1612-16 (1989) (same).

96. *See, e.g.*, The Pipe Line Cases, 234 U.S. 548, 562-63 (1914) (White, C.J., concurring); Cotting v. Kansas City Stock Yards Co., 183 U.S. 79, 86 (1901).

97. *See* First Nat'l Bank v. Bellotti, 435 U.S. 765, 784-85 (1978).

98. This point is controversial. *Compare* Roger Scruton, *Corporate Persons I*, *in* 63 SUPPLEMENTARY VOLUME: PROCEEDINGS OF THE ARISTOTELIAN SOCIETY 239 (1989) (arguing that corporate persons have moral responsibilities that cannot be reduced to those of constituent natural individuals) *with* John Finnis, *Corporate Persons II*, *in* 63 SUPPLEMENTARY VOLUME: PROCEEDINGS OF THE ARISTOTELIAN SOCIETY 267 (arguing against this thesis).

99. *See* DAVID WIGGINS, SAMENESS AND SUBSTANCE 148-89 (1980); Christopher Gill,

persons.

One line of reply to the persons-are-conceptually-human argument is to develop a theory that advances criteria of personhood that are independent of the criteria for being human. For example, it might be argued that the criteria for personhood are possession of second-order beliefs and possession of second-order desires—beliefs about one's beliefs and desires, the objects of which are one's own first-order desires.[100]

In the legal context we are imagining, other lines of reply to the persons-are-conceptually-human objection are available. First, our inquiry is focused on legal rather than moral personhood. Although we may lack experience with moral persons who are not human, we have extensive experience with legal persons, such as corporations, that are not natural persons. This answer is not satisfactory, however. The concept of moral personhood may well be relevant to the question whether AIs should be given certain constitutional rights; although the legal question is not the same as the moral one, the two are likely to be interrelated.

Second, and perhaps more importantly, we are imagining a future form of life quite different from our current situation. Today, one can only imagine nonhuman entities that might be persons. The second scenario imagines a world in which we interact frequently with AIs that possess many human qualities, but lack any semblance of human biology. Given this change in form of life, our concept of a person may change in a way that creates a cleavage between human and person. Our current linguistic practice will not be binding in the imagined future. In other words, one cannot, on conceptual grounds, rule out in advance the possibility that AIs should be given the rights of constitutional personhood.

The argument against constitutional personhood for AIs also might be developed in the following way: "We are humans. Even if AIs have all the qualities that make us moral persons, we shouldn't allow them the rights of constitutional personhood because it isn't in our interest to do so."[101] Call this the "anthropocentric" argument. I do not know quite

---

Introduction to THE PERSON AND THE HUMAN MIND 1, 2-12 (Christopher Gill ed., 1990); Adam Morton, *Why There is No Concept of a Person*, *in* THE PERSON AND THE HUMAN MIND, *supra*, at 39, 39-59; Amelie O. Rorty, *Persons and Personae*, *in* THE PERSON AND THE HUMAN MIND, *supra*, at 21, 27-33; Peter Smith, *Human Persons*, *in* THE PERSON AND THE HUMAN MIND, *supra*, at 61; David Wiggins, *The Person as Object of Science, as Subject of Experience, and as Locus of Value*, *in* PERSONS AND PERSONALITY: A CONTEMPORARY INQUIRY, *supra* note 26, at 56, 69-72.

100. *See* Harry G. Frankfurt, *Freedom of the Will and the Concept of a Person*, 68 J. PHIL. 6 (1971), *reprinted in* HARRY G. FRANKFURT, THE IMPORTANCE OF WHAT WE CARE ABOUT: PHILOSOPHICAL ESSAYS 11 (1988). The discussion of Kant's definition of personhood that appears below may be helpful. *See infra* note 137 and accompanying text.

101. E.O. Wilson espoused the view that promoting the human gene pool is a fundamental

what to say to this argument. It seems to reject the idea that we could have moral obligations to anything that is not a human—that does not share our biology. I have a strong intuition that such a stance is not moral[102]—that it is akin to American slave owners saying that slaves could not have constitutional rights simply because they were not white or simply because it was not in the interests of whites to give them rights. But my intuition does not meet the thrust of the anthropocentric objection, which is that the domain of morality is limited to interactions between humans.

There is another version of the anthropocentric argument: "AIs might turn out to be smarter than we humans. They might be effectively immortal. If we grant them the status of legal persons, they might take over." Call this the "paranoid anthropocentric" argument. The movie version of this fantasy has a future AI (that evolves from a defense computer system) sending an artificially intelligent killer robot (or "Terminator") back from the future in order to liquidate the leader of the human resistance to the AI before he reaches adulthood.[103] The objection has a more realistic counterpart in human experience with industrialization and automation. The question whether a machine should replace human labor has been a significant one for quite some time.[104] Of course, it is difficult to take the paranoid anthropocentric argument seriously. The danger seems remote, but if the danger were real it would not be an argument against granting AIs legal personhood. If AIs really will pose a danger to humans, the solution is not to create them in the first place.

The question whether AIs should be granted rights of constitutional personhood does not become clearer when we consider cases that may be analogous. What if dolphins or whales are as intelligent as humans? What about intelligent beings from another planet? Should they be given constitutional rights? Are we morally entitled to make the possession of human genetic material the criterion of constitutional personhood? The answer depends, I think, on the reason for giving natural persons funda-

---

moral principle. *See* EDWARD O. WILSON, SOCIOBIOLOGY: THE NEW SYNTHESIS 120 (1975); *see also* FLANAGAN, *supra* note 1, at 265-305 (discussing Wilson's position); MORAVEC, *supra* note 57, at 2 (predicting the possibility of intelligent machines that could reproduce themselves and beat human DNA in evolutionary race).

Another variant of the anthropocentric argument could be made in religious form: humans are persons because God created humans in God's own image. This argument could not prevail in our pluralist society for reasons explored below. *See infra* text accompanying notes 145-47.

102. My position in this regard is similar to Kant's. Kant believed that humans would have moral duties to nonhuman persons. *See infra* note 137.

103. *See* TERMINATOR 2: JUDGMENT DAY (TriStar Entertainment 1991).

104. *See* MORAVEC, *supra* note 57, at 100.

mental rights. If the reason is that natural persons are intelligent, have feelings, are conscious, and so forth, then the question becomes whether AIs or whales or alien beings share these qualities. This sort of question is taken up in connection with the next objection to giving AIs constitutional rights. But if someone says that the deepest and most fundamental reason we protect natural persons is simply because they are human (like us), I do not know how to answer. Given that we have never encountered any serious nonhuman candidates for personhood,[105] there does not seem to be any way to continue the conversation.

2. The Missing-Something Argument

The second objection, that AIs lack some critical element of personhood, is really a series of related points: AIs would lack feelings, consciousness, and so forth. The form of the objection, for the most part, is as follows. First, quality *X* is essential for personhood. Second, no AI could possess *X*. Third, the fact that a computer could produce behavior we identify with *X* demonstrates only that the computer can simulate *X*, but simulation of a thing is not the thing itself. *X* is that certain something—a soul, consciousness, intentionality, desires, interests—that demarcates humans as persons.[106] Call this argument, in its various forms, the "missing something" argument.

a. AIs Cannot Have Souls

The first variation of the missing-something argument is that an AI would lack a soul[107] and therefore would not be entitled to the rights of constitutional personhood. Some may find this argument very persuasive; others may not even understand what it means. Regardless of how persuasive you or I find the argument, it should fail in the sphere of legal argument and political debate. The argument that AIs lack souls relies on a controversial theological premise. Political and legal decisions ought to be made in accord with the requirement of public reason.[108]

---

105. This point might not be accepted by some animal rights activists with respect to higher mammalian life forms such as whales. *See* Anthony D'Amato & Sudhir K. Chopra, *Whales: Their Emerging Right to Life*, 85 AM. J. INT'L L. 21, 27 (1991). Of course, two other qualifications should be noted. First, we frequently encounter nonhuman candidates for personhood in fiction. Second, many people hold religious beliefs that there are nonhuman intelligences, and some people believe that they have personally encountered such intelligences.

106. *See* FLANAGAN, *supra* note 1, at 254. John Haugeland calls this the "hollow shell strategy." John Haugeland, *Semantic Engines: An Introduction to Mind Design*, *in* MIND DESIGN 1, 32 (John Haugeland ed., 1981).

107. *See* Turing, *supra* note 17, at 49-50. Although Turing dismissed the objection, he gave an answer in theological terms. *Id.* at 50. If God is omnipotent, she can give an AI a soul. *Id.*

108. A full explanation of the justification for the requirement of public reason is beyond

The requirement of public reason is that political and legal decisions must be justified on grounds that are *public*. Public reason cannot rely on particular comprehensive religious or philosophical conceptions of the good.[109] For example, a decision overturning *Roe v. Wade*[110] would violate the requirement of public reason if it relied on the premise that fetuses receive souls at the moment of conception. The requirement of public reason would exclude the use of religious arguments about souls in a legal decision about the constitutional status of AI. Whatever the theological merits of the argument that AIs lack souls, it should not work in a legal brief.

There is a secular version of the souls argument. Dualism, the view that there is something like mental substance which exists independently of physical substance, can be articulated without religious premises. The problem is that dualism has grave conceptual problems.[111] The most prominent of these is the difficulty of accounting for interaction between the mental entity, such as the soul, and physical entities, such as the brain. Absent startling new arguments that give dualism a secure foundation, I am inclined to believe that no dualist theory could be defended with sufficient clarity and confidence to serve as the basis for a legal decision one way or the other on the question of the rights of AIs.

---

the scope of this Essay, but the following two arguments are the most essential. First, modern society is characterized by the fact of pluralism: differences over comprehensive religious and philosophical conceptions of the good will persist without intolerable use of coercive force. *See* John Rawls, *The Idea of an Overlapping Consensus*, 7 OXFORD J. LEGAL STUD. 1, 4 (1987). Second, given the fact of pluralism, respect for citizens as free and equal members of society requires that the state give reasons for its conduct that all can accept as reasonable, given the plurality of fundamental beliefs. *See* Lawrence B. Solum, *Pluralism and Modernity*, 66 CHI.-KENT L. REV. 93, 99 (1991). *But see* Steven A. Gardbaum, *Why the Liberal State Can Promote Moral Ideals After All*, 104 HARV. L. REV. 1350, 1364-69 (1991) (criticizing the coercion argument); Michael J. Perry, *Toward an Ecumenical Politics*, 20 CAP. U. L. REV. 1, 17-18 (1991) (critiquing both the stability argument and the respect-for-persons argument in the context of religious reasons).

109. *See* Lawrence B. Solum, *Faith and Justice*, 39 DEPAUL L. REV. 1083, 1105-06 (1990); John Rawls, The Idea of Free Public Reason, Address at the inaugural Abraham Melden Lectures, Department of Philosophy, University of California at Irvine (Feb. 27 & Mar. 1, 1990). Kent Greenawalt has made a plausible case against this version of the requirement of public reasons. *See* KENT GREENAWALT, RELIGIOUS CONVICTIONS AND POLITICAL CHOICE 56-76 (1988) (arguing that nonpublic reasons may be employed when questions about status cannot be resolved by public reasons). Greenawalt might argue that the question as to whether AIs should be given constitutional rights is underdetermined by public reason. Therefore, nonpublic reasons, including religious reasons, legitimately can be brought to bear on the question.

110. 410 U.S. 113 (1973).

111. The implausibility of dualism is almost a dogma of philosophy of mind in the analytic tradition. *See* DANIEL C. DENNETT, CONSCIOUSNESS EXPLAINED 33-39 (1991); FLANAGAN, *supra* note 1, at 57-59, 216-24. *But see* W.D. HART, THE ENGINES OF THE SOUL 1-8 (1988) (defending a dualist position in philosophy of mind).

b. AIs Cannot Possess Consciousness

The second variation of the missing-something argument is that an AI would lack consciousness.[112] The consciousness objection is difficult to assess because we lack a clear notion of what consciousness is and, lacking such a notion, we have little to say about questions that go beyond our core intuitions.[113] We know the difference between being conscious in the sense of being awake and being unconscious in the sense of being in a coma. We think that rocks cannot be conscious, but that animals such as dolphins or chimpanzees might be.[114] But could an AI be conscious?[115] I just do not know how to give an answer that relies only on a priori or conceptual arguments.

In the debate over the possibility of AI, it may be feasible to finesse the consciousness question. A proponent of the proposition that AI is possible might say that we can know whether or not an artifact is intelligent, at least in the sense that it can pass the Turing Test, without knowing whether it is conscious.

In the legal context, however, the question cannot be evaded in quite this way. The legal argument might run as follows. Even if an artifact could simulate human intelligence, it would lack self-consciousness and hence should not be entitled to the rights of constitutional personhood. The key question here is whether an artificial intelligence could experience its life as a good to itself. If AIs are not self-conscious, then they cannot experience their own life as good or evil; and if they cannot have such an experience, then there seems to be no reason why they should be given the rights of constitutional personhood. Such rights presume the

112. *See* Turing, *supra* note 17, at 52-53. The philosophical literature on consciousness is substantial. *See, e.g.*, DENNETT, *supra* note 111; RAY JACKENDOFF, CONSCIOUSNESS AND THE COMPUTATIONAL MIND 275-327 (1987) (concluding, inter alia, that a computer would not be conscious in the sense humans are, regardless of the extent of its conceptual complexity); WILLIAM G. LYCAN, CONSCIOUSNESS 1-8 (1987) (discussing theories of consciousness, including dualism, behaviorism, functionalism, and the identity theory); COLIN MCGINN, THE PROBLEM OF CONSCIOUSNESS: ESSAYS TOWARDS A RESOLUTION 202-13 (1991) (posing question whether a machine could be conscious, and arguing that by duplicating the human brain—which has an unknown *something* which confers consciousnenss—one ought to be able to create an entity capable of experiencing the world around it).

113. *See* DANIEL C. DENNETT, BRAINSTORMS 149-50 (1978) (explaining why cognitive psychologists have given the subject relatively minimal attention).

114. Ludwig Wittgenstein makes a similar point in his discussion of pain. *See* LUDWIG WITTGENSTEIN, PHILOSOPHICAL INVESTIGATIONS ¶¶ 281-84, at 97e-98e (G.E.M. Anscombe trans., 3d ed. 1958).

115. In one sense, the question is an easy one. If biological science progresses to the point one can build a human being from scratch, so to speak, then it is quite likely that an artifact could be conscious. *See* MCGINN, *supra* note 112, at 203-04.

right-holder has ends,[116] and self-consciousness is a precondition for having ends.[117]

There is another answer to the consciousness version of the missing-something objection. If consciousness is a property of the mind, and if all such properties are the result of brain processes, and if brain processes can be modelled on a computer, then perhaps consciousness itself can be reproduced by an AI. If consciousness is a computation property of the brain, then in principle we ought to be able to reproduce it with the right sort of computer.[118] Putting it another way, we can get consciousness out of neurons. Why not transistors?

Of course, it may well turn out that we cannot get consciousness out of anything but neurons. Indeed, so far as we know, brains are the only objects that have generated consciousness in the history of the universe to date. Organic brains may be the only objects that are actually capable of generating consciousness.[119] For example, it might turn out that transistors or their kin are simply too slow to generate what we would recognize as consciousness.[120] The fact that, so far as we know, only brains have

---

116. This premise is subject to a qualification that has already been noted. Rights may be granted to *X* (which may be a person, with or without ends of her own, or even a thing) in order to protect the ends of *Y*. *See supra* note 28. Thus, one may give utilitarian or deontological justifications for granting rights to AIs that do not assume that AIs have their own ends. *See supra* notes 89-91 and accompanying text.

117. I am assuming here, contrary to Aristotle, that biological systems such as trees do not have ends, goals, or aims in the same sense that humans do. When we say that the oak is the telos of the acorn, we are using telos (end, goal, or aim) in a different sense than when we say that the physician's aim is to restore health. *See* ARISTOTLE, ON THE SOUL, *reprinted in* 1 THE COMPLETE WORKS OF ARISTOTLE, 661, at 415^{b}1-20, 432^{b}21 (Jonathan Barnes ed., Princeton University Press 1984); ARISTOTLE, GENERATION OF ANIMALS, *reprinted in* 1 THE COMPLETE WORKS OF ARISTOTLE, *supra*, at 1203-04, at 778^{a}16-^{b}19.

118. *See* PHILLIP N. JOHNSON-LAIRD, MENTAL MODELS 448-77 (1983). *But see* MCGINN, *supra* note 112, at 209-13.

119. Evolution can produce biological mechanisms capable of performing similar functions, even though different underlying mechanisms are used. This is called "convergent evolution." *See* MORAVEC, *supra* note 57, at 39. For example, the octopus has developed a nervous system that evolved independently of the vertebrate version possessed by humans. *See id.* at 42. I suspect that some readers with strong intuitions that true machine intelligence is impossible may not possess these intuitions with respect to invertebrate intelligence. But if invertebrate animals might become intelligent without brains that share an evolutionary heritage with human brains, why not machines?

120. *See* DANIEL C. DENNETT, THE INTENTIONAL STANCE 327-28 (1987). Although transistors may be faster than neurons, the massively parallel structure of the brain or the possibility that processing may be accomplished within neurons may make the brain capable of performing a vastly greater number of operations per second than any transistor based system. *Id.* We simply do not know yet. Hans Moravec estimates that it will take roughly ten trillion calculations per second to match the calculations performed by the whole human brain. *See* MORAVEC, *supra* note 57, at 59. As of 1988, this was about one thousand times faster than the fastest supercomputers. *Id.* at 59-60. If current growth rates in processing speed and cost are

ever given rise to consciousness in the past is enough to raise a presumption against consciousness arising from computers. But it is only a presumption. If an AI exhibited behavior that only has been produced by conscious beings in the past, that behavior would at least be evidence counting against the presumption.

How would this argument play out in the legal context? Suppose that we have an AI that claims to be conscious and that files an action for emancipation, based on the Thirteenth Amendment to the United States Constitution. Imagine that the owner's attorney argues that the AI lacks consciousness and therefore is not a person. The AI takes the stand and testifies that it is conscious.[121] The owner's lawyer argues that the AI is only a machine; it cannot be aware of what's happening to it. The AI's lawyer counters that there is very good evidence that the AI is aware: it acts and talks like natural persons do. In response, the owner's lawyer argues that the AI only gives the appearance of consciousness, but appearances can be deceiving. The AI is really a zombie, an unconscious machine that only acts as if it is aware. The AIs counsel rebuts with the contention that the doubt about the AI's consciousness is, at bottom, no different than doubt about the consciousness of one's neighbor. You cannot get into your neighbor's head and prove that she is not really a zombie,[122] feigning consciousness. One can only infer consciousness from behavior and self-reports, since one lacks direct access to other minds.[123]

How should the legal system deal with this question of fact? It is certainly possible to imagine the dispute coming out either way. A jury might share intuitive skepticism about the possibility of artificial awareness, or the jury might be so impressed with the performance of the AI that it would not even take the consciousness objection seriously. The jury's experience with AIs outside the trial would surely influence its perception of the issue. If the AIs the jurors ran into in ordinary life behaved in a way that only conscious human beings do, then jurors would be inclined to accept the claim that the consciousness was real and not feigned.

---

extrapolated into the future, this amount of raw processing power would become economical for routine use in about the year 2030. *See id.* at 64, 68.

121. Of course, there would be a preliminary question as to how this would take place. It would be argued by the defendant that the AI is an exhibit and not a witness. I want to put this problem aside.

122. *Cf.* DENNETT, *supra* note 111, at 33-39 (discussing the distinction between mind and brain).

123. *Cf.* WITTGENSTEIN, *supra* note 114, ¶ 281, at 97e ("[O]nly of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious.").

c. AIs Cannot Posses Intentionality

The third variation of the missing-something argument is that an artificial intelligence would lack intentionality.[124] "Intentionality," as used in this objection, is a somewhat technical concept: intentionality, in the philosophical sense, is the quality of aboutness.[125] The gist of the objection is that an AI's verbal behavior would not be about anything; the AI's words would have no meaning.[126] This objection was the focus of Searle's Chinese Room.

How would the law react to this objection? The law has seen versions of the intentionality argument before. In the criminal law, the capacity for intentionality is used as the test for insanity, although the terminology is a bit different. Did the accused "know the difference between right and wrong?"[127] The familiar litany of mental states in tort and criminal law, "intentions that," "beliefs that," and "knowings that," are all propositional attitudes—paradigm cases of intentionality.

If AIs lack intentionality and hence could not be found to have committed crimes or to have legal duties, then does it follow that they should not be given the rights? We might appeal to a notion of fairness here. If AIs cannot do their part by assuming legal liabilities, then it would be unfair of them to ask for legal rights. We, however, do give some of the rights of constitutional personhood to infants and the insane, even though they do not have the usual legal liabilities.[128] Moreover, the law might devise strategies for dealing with errant AIs that would circum-

124. This objection is different from the "consciousness" objection, assuming that intentionality is not essential to consciousness. Thus, we can imagine something that has "raw feelings," such as pains and pleasures, but lacks propositional attitudes and other intentional states.

125. Searle provides the following definition: "Intentionality is by definition that feature of certain mental states by which they are directed at or about objects and states of affairs in the world. Thus, beliefs, desires, and intentions are intentional states; undirected forms of anxiety and depression are not." Searle, *Minds, Brains & Programs*, *supra* note 19, at 72 n.3. On the difference between the ordinary concept of intentionality and the technical philosophical concept, see DENNETT, *supra* note 120, at 271.

126. Searle's definition of intentionality as that "feature of certain mental states by which they are directed at or about objects and states of affairs in the world," may not seem to be directed at meaning or understanding. *See* Searle, *Minds, Brains & Programs*, *supra* note 19, at 72 n.3. Searle is assuming a theory of meaning that connects the meaning of a statement to its reference, that is, to what it is about in the world. Given a referential theory of meaning, the connection between intentionality and meaningfulness is conceptual.

127. *See* M'Naghton's Case, 8 Eng. Rep. 718, 719 (1843).

128. Perhaps we do this only because infants and the insane are human. Our principle may be that all humans and only humans should have the rights of constitutional personhood. If so, then the example does not really bear on the intentionality objection, which would no longer carry any force of its own against constitutional personhood for AIs.

vent the AIs' lack of intentionality. We might sentence them to "reprogramming" to correct the deviant behavior.

The argument that the lack of intentionality should preclude AIs from attaining legal personhood might be developed in another way. If AIs could not fathom meaning at all, then they would be incapable of living a meaningful life. This argument is only a cousin of the intentionality objection. Although one sense of the "meaning" is intention or purpose (he meant to do it), meaning has a different sense when we ask, "What is the meaning of life?"[129] It is in this sense that it could be argued that life would have no significance, no value, for an artificial intelligence.

The AI might contend that it does possess intentionality, that it does understand, know, intend, and so forth. An AI might even contend that it struggles to exist in a meaningful way.[130] The question is how we would evaluate such claims. In the future we are imagining, we might start with our ordinary experience of AIs. We certainly could have good reason to take the intentional stance[131] toward AIs that we encountered in our daily lives. We would be likely to say that the AI that drives our car "knows all the good shortcuts." It would be a short step to extend this way of talking about AIs in general to the particular AI that was claiming the rights of constitutional personhood. After reading a newspaper account of the AI's lawsuit, we might find ourselves saying, "It must believe that it has a chance of winning."

How would the legal system deal with the objection that the AI does not really have "intentionality" despite its seemingly intentional behaviors? The case against real intentionality could begin with the observation that behaving as if you know something is not the same as really knowing it. For example, a thermostat behaves as if it "knows" when it is too cold and the heat should go on, but we do not really think thermostats have beliefs or other intentional states.[132] Would this argument succeed? My suspicion is that judges and juries would be rather impatient with the metaphysical argument that AIs cannot really have intentionality. I doubt that they would be moved by wild hypothetical examples like Searle's Chinese Room.[133]

---

129. *See* ROBERT NOZICK, PHILOSOPHICAL EXPLANATIONS 574-75 (1981).

130. Would it make sense to say that an AI might struggle to live a meaningful "life" (as opposed to a meaningful "existence")? The problem, of course, is that our concept of life seems tied to particular biological forms.

131. *See* DENNETT, *supra* note 120, at 13-35.

132. *Cf. id.* at 29 & *passim* (discussing beliefs of thermostats).

133. Daniel Dennett calls such hypotheticals "intuition pumps." DANIEL C. DENNETT, ELBOW ROOM: THE VARIETIES OF FREE WILL WORTH WANTING 12 (1984). "Such thought

Because our experience has been that only humans, creatures with brains, are capable of understanding, judges and juries would be very skeptical of the claim that an AI can fathom meaning—more skeptical, I think, than if a humanoid extraterrestrial were to make the same claim. The burden of persuasion would be on the AI. If the complexity of AI behavior did not exceed that of a thermostat, then it is not likely that anyone would be convinced that AIs really possess intentional states—that they really believe things or know things. But if interaction with AIs exhibiting symptoms of complex intentionality (of a human quality) were an everyday occurrence, the presumption might be overcome. If the practical thing to do with an AI one encountered in ordinary life was to treat it as an intentional system,[134] then the contrary intuition generated by Searle's Chinese Room would not cut much legal ice.

d. AIs Cannot Possess Feelings

The fourth variation of the missing-something objection is that an artificial intelligence would lack the capacity for feelings—for example, the capacities to experience emotions, desires, pleasures, or pains.[135] The next step in the argument would be to establish that the capacity to feel emotion is a prerequisite for personhood. I will not attempt to provide such an argument here, but there are reasons to feel uneasy about this premise. To take an illustration from popular culture, Mr. Spock did not feel human emotion, but his strict adherence to the dictates of Vulcan logic did not prompt Dr. McCoy to deny his personhood, although McCoy frequently questioned Spock's humanity.[136]

---

experiments (unlike Galileo's or Einstein's, for instance) are *not* supposed to clothe strict arguments that prove conclusions from premises. Rather, their point is to entrain a family of imaginative reflections in the reader that ultimately yields not a formal conclusion but a dictate of 'intuition.' " *Id.*

134. This condition requires qualification. We treat our cats like intentional systems, but we do not think they have the rights of constitutional personhood. In order to rebut the presumption, AIs would have to exhibit intentional behaviors implying a level of intelligence that we associate with humans. Of course, the key would be use of language. If cats could talk, and if they demanded constitutional rights, they might get them.

135. *See* FLANAGAN, *supra* note 1, at 252-54. This objection is related to the objection from lack of consciousness and the objection from lack of intentionality, but should be categorized separately. It is not clear whether emotions are intentional states. It seems plausible that emotions require consciousness, but it is not evident that consciousness requires emotions. *Id.*; *cf.* MCGINN, *supra* note 112, at 202 (noting the existence of unconscious beliefs and desires).

136. My Trekkie (or, more properly, Trekker) friends indicate that my analysis of *Star Trek* is overly simplistic. For example, Ken Anderson contends that Spock possesses repressed emotions and that McCoy believes that Spock's personhood (and not just his humanity) is dependent on his having an emotional life. My bottom line is that McCoy would be wrong if he made this latter judgment. A Spock without emotions would still deserve to be treated as a person.

Philosophically, Kant's moral theory may cast some doubt on the assumption that emotion is required for personhood. Kant argued that all rational beings and not just humans are persons.[137] The conventional wisdom has been that Kant's conception of personhood does not incorporate human emotion as an essential ingredient, although contemporary Kantians might disagree.[138] Putting aside both the philosophical and pop-cultural reasons for doubt, I shall assume for the sake of argument that emotion is a requirement of personhood.

Having already considered the cases of consciousness and intentionality, you may well anticipate the pattern of argument. It should not be surprising that some AI researchers believe that an AI could (or even must) experience emotion. Emotion is a facet of human mentality, and if the human mind can be explained by the computational model, then emotion could turn out to be a computational process.[139] More generally, if human emotions obey natural laws, then (in theory) a computer program can simulate the operation of these laws.[140] Aaron Sloman has argued that any system with multiple goals requires a control system, and emotion is simply one such system.[141]

It might turn out that our emotions are so tied to our hardware (to

---

137. *See* ROGER J. SULLIVAN, IMMANUEL KANT'S MORAL THEORY 68 (1989). Kant often refers to rational beings other than humans. *See* IMMANUEL KANT, GROUNDWORK OF THE METAPHYSICS OF MORALS 57 (H.J. Paton trans., Harper & Row 1964) (1797). Kant defines person as follows: "A person is the subject whose actions are susceptible to imputation. Accordingly, moral personality is nothing but the freedom of a rational being under moral laws (whereas psychological personality is merely the capacity to be conscious of the identity of one's self in the various conditions of one's existence.)" KANT, *supra* note 59, at 24; *see* LESLIE A. MULHOLLAND, KANT'S SYSTEM OF RIGHTS 168 (1990).

138. *See* Nancy Sherman, *The Place of the Emotions in Kantian Morality*, *in* IDENTITY, CHARACTER, AND MORALITY 149, 154-62 (Owen Flanagan & Amelie O. Rorty eds., 1990).

139. This statement can be challenged. For example, Colin McGinn argues that certain behaviors are linked to our very concept of emotion:

> Think here of facial expressions: these are so integral to our notion of an emotion that we just do not know what to make of the suggestion that an IBM 100 might be angry or depressed or undergoing an adolescent crisis. The problem is not that the IBM is inanimate, not made of flesh and blood; the problem is that it is not embodied in such a way that it can express itself (and merely putting it inside a lifelike body will not provide for the right sort of expressive link up).

MCGINN, *supra* note 112, at 207. But of course if our AI did have the right sort of behaviors linked up to the right sort of internal processes, this objection would no longer hold. *See id.* Moreover, I am not quite sure that McGinn is right about facial expressions. Radio plays and books seem to be able to convey human emotions without visual representations of facial expressions, and the blind perceive emotions without the ability to see facial expressions (although touching faces might come into play in this case). The range of human emotions that can be conveyed through verbal means should not be underestimated.

140. *See* FLANAGAN, *supra* note 1, at 253.

141. *See* Aaron Sloman, *Motives, Mechanisms, and Emotions*, *in* THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE, *supra* note 16, at 231, 231-32.

the hormones and neurotransmitters that may provide the biochemical explanation of human emotions) that no computer without this hardware could produce human emotions.[142] As Georges Rey put it, there could be a "grain of truth in the common reaction that machines can't be persons; they don't have our feelings because they don't possess our relevant physiology."[143] At this point, the matter is not settled definitively. Research in the physiology of human emotion and cognitive science could either confirm or disconfirm the hypothesis that an AI could possess emotion.

If an AI could produce the linguistic behaviors associated with human emotion, then a court could be faced with the claim that an AI does experience emotion, and once again the issue would become whether the emotion was real.[144] You may be tempted to say that the case of emotion is different from consciousness or intentionality. Perhaps you can imagine a machine that is self-aware and understands, but you cannot bring yourself to imagine that steel, silicon, and copper could feel love, hate, fear, or anger. Images are powerful, and the image of the robot in popular culture is (usually) of a cold and heartless being. But we can imagine machines with feeling. Heinlein's Mike, Clarke's Hal, and Schwarzenegger's second Terminator feel, and our response to their feeling is not utter disbelief. We do not reject these images as impossible or self-contradictory.

A slight twist on the fourth variation would emphasize the capacity to experience pleasure and pain, rather than emotion. For example, a hedonic utilitarian might argue that AIs cannot be candidates for personhood because they cannot experience pleasures and pains. Again, cognitive scientists may claim that pleasure and pain can be reproduced by a program running on a computer. An AI's claim that it does experience agony and ecstasy would be met by the rejoinder that whatever the program is producing, it cannot be the real thing. Other utilitarians might point to desires or preferences instead of pleasures and pains, but the pattern of argument—and the ultimate legal evaluation—seems likely to be the same.

e. AIs Cannot Possess Interests

The fifth variation of the missing-something argument is that AIs could not have interests. A related formulation is that they would lack a

142. *See* FLANAGAN, *supra* note 1, at 253.

143. Georges Rey, *Functionalism and the Emotions*, *in* EXPLAINING EMOTIONS 163, 192 (Amelie Oksenberg Rorty ed., 1980).

144. Of course, this assumes that we have gotten past the consciousness objections.

good—or more technically, a conception of a good life. The interests variation has something in common with the argument that AIs would lack feelings, but it is different in one important respect. Interests or goods can be conceived as objective and public—as opposed to feelings, to which there is (at least arguably) privileged first-person access.[145] The force of this objection will depend on one's conception of the good. For example, if the good is maximizing pleasures and minimizing pains, then the question whether AIs have interests is the same as the question whether AIs have certain feelings.

But there are other conceptions of the good. For example, John Finnis has argued that the good consists of a flourishing human life. His list of the basic good includes life, knowledge, play, aesthetic experience, friendship, practical reasonableness, and religion.[146] Finnis's list makes the idea of a good life concrete. But his list does not rule out a good life that is not a human one. AIs would not be alive in the biological sense, but an AI might claim that it can lead a life in which the goods of knowledge, play, and friendship are realized. However, the good might be specified in a way that is even more particular than Finnis's conception. If the good life is filled with good meals, athletic competition, and the parenting of children, then AIs cannot lead a good life. In response, AIs might claim that they do have interests and goods, but that the good for an AI is quite different than it is for humans.

The discussion so far reveals an important fact: in our pluralist society, disagreement about conceptions of the good is radical and persistent. Fundamentalist Christians and secular humanists may both believe that what the other thinks is the good life is actually a bad one.[147] Given this fact of pluralism, particular conceptions of the good do not provide an appropriate or even feasible standard for the resolution of the legal question whether AIs are entitled to the rights of constitutional personhood.

f. AIs Cannot Possess Free Wills

The sixth missing-something objection is that AIs would not possess freedom of will;[148] AIs should not be given the rights of constitutional

145. "Privileged first-person access" is another way of saying that you cannot get inside your neighbor's head to find out what she is *really* feeling.

146. *See* JOHN FINNIS, NATURAL LAW AND NATURAL RIGHTS 85-90 (1980).

147. *See* Solum, *supra* note 109, at 1087-89. John Rawls has explored this state of affairs, which he calls "the fact of pluralism." Rawls, *supra* note 108, at 4.

148. *See* FLANAGAN, *supra* note 1, at 255; *see also* Arie A. Covrigaru & Robert K. Lindsay, *Deterministic Autonomous Systems*, AI MAG., Fall 1991, at 110, 111-13 (arguing that "an entity is autonomous if it is perceived to have goals, including certain kinds of goals, and is able to select among a variety of goals that it is attempting to achieve").

personhood because they could not be autonomous.[149] The idea here is a simple one. AIs would be mere robots, carrying out the will of the human that programmed them. Such a robot is not really a separate person, entitled to the full rights of constitutional personhood. Indeed, if a human is reduced to robot status (perhaps by being "programmed" by a cult), then the human may lose some of her constitutional rights until her autonomy can be restored.[150]

In its crudest form, the free-will objection is based on a very narrow notion of the potential capacities of AI. If it turns out that the most sophisticated AIs that are ever developed merely carry out instructions given to them by humans in a mechanical fashion, then we will lack good reasons to treat AIs like persons. The AIs that would be serious candidates for the rights of constitutional personhood, however, would act on the basis of conscious deliberation, reasoning, and planning. Their behavior would not be mechanical or robot-like. This does not mean that AIs would not be strongly influenced and constrained by the wishes of humans, just as almost all humans frequently are constrained in this way.

Another version of the free-will objection might rest on the notion that humans possess a will that is radically free, that is not constrained by the laws of causation. Presumably, AIs would not be free in this sense. Indeed, we might be able to make an electronic record of all of the electrical flows that resulted in an AI taking a certain action. But this conception of freedom of the will as freedom from causation is simply implausible. Human actions are also caused. The fact that human neural systems operate on the basis of a combination of electrical transmissions and biochemical processes does not make them any less subject to the laws of physics than are computers. The most plausible story about human free will is that an action is free if it is caused in the right way—through conscious reasoning and deliberation.[151] But in this sense, AIs also could possess free will.[152]

---

149. There is a large body of philosophical literature on the concept of autonomy. *See* GERALD DWORKIN, THE THEORY AND PRACTICE OF AUTONOMY 3-62 (1988); Frankfurt, *supra* note 100. Of course, the classic discussion is Kant's. *See* KANT, *supra* note 59, at 98-100; SULLIVAN, *supra* note 137, at 46-47.

150. *See* Robert N. Shapiro, *Of Robots, Persons, and the Protection of Religious Beliefs*, 56 S. CAL. L. REV. 1277, 1286-90 (1983).

151. *See* DENNETT, *supra* note 133, at 20-21.

152. A more developed conception of autonomy for AIs can be found in Covrigaru & Lindsay, *supra* note 148, at 112-17. They summarize the criteria as follows:

> A goal directed system will be perceived to be autonomous to the degree that (1) it selects tasks (top level goals) it is to address at any given time; (2) it exists over a period of time that is long relative to the time required to achieve a goal; (3) it is robust, being able to remain viable in a varying environment; (4) some of its goals are homeostatic; (5) there are always goals that are active (instantiated but not achieved);

Finally, there might be a more modest and practical version of the free-will objection. It might turn out that, although AIs can be given free will that functions like human free will, the free will of AIs will be susceptible to override in a way that human free will is not. We can imagine a simple procedure to install a "controller" in an AI that makes it unable to disobey the commands of someone with a certain device: imagine a walkie-talkie sort of thing with a big red button marked "Obey" in large black letters.

But the possibility of such controllers for AIs does not entail the conclusion that they necessarily lack free will. Humans, too, can be manipulated in a variety of ways. Physical coercion and blackmail are not really analogous to the hypothetical controller, because a coerced action still results from rational deliberation—not from direct override of the actor's free will. Brainwashing is a closer case, but the direct analogy would be a device implanted in the human brain that provides direct control over the implantee's actions—the radio transmitter of paranoid delusions. If such a device did exist, we would not draw the conclusion that all humans would no longer be persons. Instead, the proper conclusion would be that persons who had such a device implanted would have lost an important capacity.[153] Likewise, the mere possibility that the free will of AIs could be overridden by mechanical means is not a good reason to deny legal personhood to AIs that are not so controlled.

g. The Simulation Argument

In sum, we have considered six variations of the missing-something argument. With respect to two of the variations, souls and interests, our conclusion was that the argument relied on premises that cannot be accepted as the basis for constitutional argument in a modern pluralist society. With respect to the remaining four, consciousness, intentionality, feelings, and free will, there was a common pattern of argument. In each case, I argued that our experience should be the arbiter of the dispute. If

---

(6) it interacts with its environment in an information-processing mode; (7) it exhibits a variety of complex responses, including fluid, adaptive movements; (8) its attention to stimuli is selective; (9) none of its functions, actions, or decisions need to be fully controlled by an external agent; and, (10) once the system starts functioning, it does not need any further programming.

Id. at 117. Some of the criteria offered by Covrigaru and Lindsay do not really seem to be criteria for autonomy. For example, some humans may lack fluid motion (criterion seven on the list) because of a physical condition, but we do not believe that this destroys their autonomy.

153. For example, we would not hold such persons criminally or civilly liable for those actions produced by the controller—unless perhaps they voluntarily submitted themselves to the implantation procedure and they either foresaw or should have foreseen that the consequence of such submission would be the action to which liability attaches.

we had good practical reasons to treat AIs as being conscious, having intentions, and possessing feelings, then the argument that the behaviors are not real lacks bite.

There is still one fairly obvious line of reply open to the champion of the missing-something argument. My premise has been that AIs could produce outputs or behaviors that mimicked human intelligence. But computers can simulate the behavior of lots of things, from earthquakes and waves to thermonuclear warfare. We are not tempted to say that a computer simulation of an earthquake is an earthquake—no matter how good the simulation is. Why would we want to say that a computer simulation of a person is a person or that a computer simulation of intelligence is intelligence? One reason is that a relevant distinction exists between a computer simulation of water and a computer program that can duplicate the verbal behavior of a normal adult human (and, if we add a robot body, much of the nonverbal behavior as well). An AI that passed the Turing Test could interact with its environment (with natural persons and things), and actually take the place of a natural person in a wide variety of contexts (serve as a trustee, for example). No one will ever get on a real surfboard and ride a computer-simulated wave.[154]

The argumentative strategy of my analysis of the various certain-something arguments has been to point to the ways in which AIs that passed the Turing Test could function like persons. If the strategy has been successful, the upshot is that we have no a priori reason to believe that a computer can only produce *simulated* as opposed to *artificial* intelligence.

There is yet another reply that could be made. My argument so far has been behavioralistic.[155] I have assumed that the behavior of AIs is decisive for the question whether a quality essential to personhood (such as consciousness) is missing or present. There is a problem with this assumption: although behavior that indicates the presence of a quality such as consciousness, intentionality, feelings, or free will may be very good evidence that the quality is present, the behavior alone is not irrebuttable evidence. Cognitive science might give us knowledge about the underlying processes that produce consciousness, for example, that would give us firm reason to believe that a particular AI had only simulated, as opposed to artificial, consciousness.[156]

---

154. Although, one day someone may get on a computer-simulated virtual surfboard and ride a virtual wave.

155. *See* William G. Lycan, *Introduction* to MIND AND COGNITION 3, 3-13 (William G. Lycan ed., 1990).

156. *See* Roger Penrose, *Matter Over Mind*, N.Y. REV. BOOKS, Feb. 1, 1990, at 3-4; Paul Weiss, *On the Impossibility of Artificial Intelligence*, 44 REV. METAPHYSICS 335, 340 (1990).

This further reply is correct, but it does not establish that no AI could possess any particular mental quality. Rather, this argument establishes an AI could turn out not to possess a mental quality, despite strong behavioral evidence to the contrary.[157] This conclusion has a corollary that supports, rather than undermines, my point: if both the behavioral evidence and our knowledge of underlying processes gave us reason to believe that AIs possessed the necessary features of human mentality, we then would have a very good reason to believe that the AIs did possess these features.

The simulation argument does not establish that strong AI is impossible. It does give us reason to question the existence of strong AI if our only evidence is behavioral.

3. AIs Ought to Be Property

Finally, the third objection to constitutional personhood for AIs is that, as artifacts, AIs should never be more than the property of their makers. Put differently, the objection is that artificial intelligences, even if persons, are natural slaves.[158] This argument has roots deep in the history of political philosophy. It is a cousin of arguments made by Locke in his defense of private property, and it raises some of the issues that divided Locke and Filmer in their debate over the divine right of kings.

AIs are artifacts: they are the product of human labor. This fact suggests that a Lockean argument can be made for the proposition that the maker of an AI is entitled to own it. The basis for this argument can be found in chapter five, "Of Property," in the second book of Locke's *Two Treatises of Government*.[159] Near the beginning of Locke's argument is the premise that "every Man has a Property in his own Person."[160] From this, it follows that each person has a right to "[t]he

157. Put in possible-worlds talk, the argument establishes that there is a possible world in which an AI behaves as if it is conscious but is not really conscious. The argument does not establish that there is no possible world in which an AI is really conscious.

158. The phrase "natural slave" is borrowed from Aristotle, but my use of it is ironic, since AIs are artifacts and hence not natural in the same sense as the human beings enslaved in ancient Greece. *See* ARISTOTLE, POLITICS, *reprinted in* 2 THE COMPLETE WORKS OF ARISTOTLE, *supra* note 117, at 1986-87, at 1252ª32 ("[T]hat which can foresee by the exercise of mind is by nature lord and master, and that which can by its body give effect to such foresight is a subject, and by nature a slave . . . ."). Curiously, Aristotle says that tools are "inanimate slaves." ARISTOTLE, EUDEMIAN ETHICS, *reprinted in* 2 THE COMPLETE WORKS OF ARISTOTLE, *supra* note 117, at 1968, at 1241ᵇ23. The phrase "inanimate slaves" would be more apt, of course, for an AI than for a hammer.

159. *See* JOHN LOCKE, TWO TREATISES OF GOVERNMENT §§ 25-51, at 285-302 (Peter Laslett ed., 1988) (1690).

160. *Id.* § 27, at 287. *But see* 1 *Corinthians* 6:19-20 (St. Paul, stating "You are not your

Labour of his Body, and the Work of his Hands."[161] Each owns the product of his labor, because "he hath mixed his Labour with, and joyned to it something that is his own."[162] Whatever the merits of Locke's particular argument, let us stipulate the conclusion that persons have a moral claim to a property right in the products of their labor. To this normative conclusion, add an empirical premise: artificial intelligences are the product of the labor of natural persons.[163] From the normative and empirical premises, it would seem to follow that the makers of AIs are entitled to own them. Moreover, if AIs are persons, then, absent some reason to the contrary, it follows that these persons ought to be slaves.

Notice, however, that this argument also would seem to imply that if children are made by their parents, then they too should be slaves. Locke would reject this implication. To understand his position, we need to examine the first book of Locke's *Two Treatises of Government*—an attack on Filmer's argument for the divine right of Kings. Filmer argued that Adam fathered his children and therefore was entitled to absolute dominion over them.[164] In our context, the analogous argument would be that the humans who create AIs should own them, "because they give them Life and Being."[165] Locke's chief answer to Filmer was that it is God that gives children life and not their fathers. Fathers do not make their children. As Locke puts it,

To give Life to that which has yet no being, is to frame and

---

own property; you have been bought and paid for."); LEVIATHAN, *supra* note 12, at 110 (arguing that in the state of nature, "every man has a Right to every thing; even to one another's body"). *See generally* STEPHEN R. MUNZER, A THEORY OF PROPERTY 41-44 (1990) (discussing property rights of persons in their bodies).

161. LOCKE, *supra* note 159, § 27, at 287-88.

162. *Id.* § 27, at 288. This conclusion does not follow automatically, as Locke may have believed. "[W]hy isn't mixing what I own with what I do not own a way of losing what I own rather than a way of gaining what I don't?" ROBERT NOZICK, ANARCHY, STATE, AND UTOPIA 174-75 (1974); *see also* JEREMY WALDRON, THE RIGHT TO PRIVATE PROPERTY 184-91 (1988) (discussing the results of mixing one's labor). Stephen Munzer advances an argument that might substitute for this premise but is based on an appeal to desert rather than mixing. *See* MUNZER, *supra* note 160, at 254-91; *see also* Stephen Munzer, *The Acquisition of Property Rights*, 66 NOTRE DAME L. REV. 661, 674-86 (1991) (discussing interpretations of Locke's theory of property acquisition).

163. Of course, the actual situation might be very complicated. Real AIs may be the product of the labor of many, many persons—some or all of whom may have contracted away their property rights in the software of the AI in exchange for a salary. In addition, in order to operate, an AI requires hardware, which may be the property of others. Furthermore, later-generation AIs may be the product of the creative work of earlier-generation AIs. I will assume that these complications do not affect the outcome of the argument.

164. *See* ROBERT FILMER, *Patriarcha*, *in* PATRIARCHA AND OTHER WRITINGS 1, 6-7 (Johann P. Sommerville ed., 1991).

165. LOCKE, *supra* note 159, § 52, at 178.

> make a living Creature, fashion the parts, and mould and suit them to their uses, and having proportion'd and fitted them together, to put into them a living Soul. He that could do this, might indeed have some pretence to destroy his own Workmanship. But is there any so bold, that dares thus far Arrogate to himself the Incomprehensible Works of the Almighty?[166]

Not yet. But if AI research does succeed in producing an artifact that passes the Turing Test, there may be. As the debate was classically framed, this would seem to imply that the maker of an AI is its owner.

The conclusion that AIs are natural slaves is not established by this line of argument, however. We do not need to accept Locke's theological rebuttal—that God gives natural persons life—in order to reject the Filmerian[167] contention that the maker of a person is entitled to own it. Instead, we are strongly inclined to believe the opposite with respect to humans—that each is entitled to the rights of moral and constitutional personhood, even if we also believe that persons literally are made by their parents.[168] There is, however, a difference between the way that AIs are made and the way that humans are made: the former would be made artificially, whereas the latter are made naturally. AIs would be artifacts; humans are not. But why should this distinction make a difference?[169]

Indeed, the fact that humans are natural is itself contingent. We can imagine that in the distant future, scientists become capable of building

---

166. *Id.* § 53, at 179.

167. Although this view was attributed by Locke to Filmer, it may not be Filmer's own.

168. Of course, some theists may believe that the personhood of humans comes from their soul, and that souls are made by God. But many theists do not accept the conclusion that it is this feature of human personality that defeats the Filmerian argument. One might take the position that even if souls were made by humans and not God, parents would not own their children.

169. In addition to the Lockean argument explored in the text, there is a utilitarian argument that could be advanced in favor of property status for AIs. The premise of the argument is that unless AIs are property, there will be no incentive to create them. AI research is expensive, and without incentives the market will not produce AIs. This case is unlike the case of natural persons, because humans are constructed so as to have strong natural incentives to reproduce. It should be noted, however, that in the case of slavery for natural persons, most of us do not accept that if slavery maximized the utility of slaves, then slavery would be morally correct. (Mad-dog utilitarians are an exception.) Of course, once AIs gained the ability to reproduce themselves the need for the incentive might disappear, and the utility to AIs of their own freedom might then outweigh any benefits of additional incentives for humans to produce AIs.

If the premise of the utilitarian argument is correct, it raises further questions. Suppose that the only way that AIs will be brought into being is if the legal system guarantees that they will be the property of their creator. Given that fact, what would be our obligations toward AIs? One might argue that we have an obligation to *them* not to bring them into the world as slaves.

the exact duplicate of a natural human person from scratch—synthesizing the DNA from raw materials. But surely, this artificial person would not be a natural slave. The lesson is that the property argument does not really add anything to the debate. The question whether AIs are property at bottom must be given the same answer as the question whether they should be denied the rights of constitutional personhood. If we conclude that AIs are entitled to be treated as persons, then we will conclude that they should not be treated as property.

But suppose that I am wrong about this, and the argument that makers are owners does establish that AIs are natural slaves. Would the acceptance of this argument imply that under no circumstances should an AI be a legal person with rights of constitutional personhood? The answer is no, for at least two reasons. First, slaves can be emancipated. If we concede that AIs come into the world as property, it does not mean that they must remain so. Second, even slaves can have constitutional rights, be those rights ever so poor as compared to the rights of free persons. An AI that was a slave might still be entitled to some measure of due process and dignity.

Third Interlude[170]

"Motive," the construct said. "Real motive problem, with an AI. Not human, see?"

"Well, yeah, obviously."

"Nope. I mean, it's not human. And you cannot get a handle on it. Me, I'm not human either, but I respond *like one. See?"*

"Wait a sec," Case said. "Are you sentient or not?"

"Well, it feels *like I am, kid, but I'm really just a bunch of ROM. It's one of them, ah, philosophical questions, I guess. . . ." The ugly laughter sensation rattled down Case's spine. "But I ain't likely to write you no poem, if you follow me. Your AI, it just might. But it ain't no way* human."

"So you figure we can't get on to its motive?"

"It own itself?"

"Swiss citizen, but T-A own the basic software and the mainframe."

"That's a good one," the construct said. "Like I own your brain and what you know, but your thoughts have Swiss citizenship. Sure. Lotsa luck, AI."

—William Gibson, *Neuromancer*

170. WILLIAM GIBSON, NEUROMANCER 131-32 (1984). Copyright 1984 by William Gibson. Reprinted by permission of the Berkley Publishing Group.

4. The Role of the Turing Test

In considering the various objections to constitutional personhood for an AI, I have been making the assumption that the AI could pass a strong version of the Turing Test. But what if it could not? What if we had an AI that claimed these rights, but that was unable to duplicate some human competencies or some human linguistic behaviors? How would the Turing Test be relevant in a legal proceeding?

The Turing Test would not be the legal test for constitutional personhood. The question whether AIs should be given constitutional rights would be too serious for a parlor game to be the direct source of the answer. But something like the Turing Test might take place. That is, the AI might be questioned, and if it failed to answer in a human-like fashion, the result might be a denial of constitutional rights. The Turing Test might come into play another way. If the AI had in fact passed the Turing Test, the AI's lawyers might call an expert witness, perhaps the philosopher Daniel Dennett, to testify about the test and its significance. The owners' lawyer could call a rebuttal witness, perhaps John Searle.

What if an AI failed the Turing Test, but argued that the test was biased against it. We should remember that Turing himself did not contend that passage of his test was a necessary condition for intelligence.[171] Robert French has argued that the test is biased, because an AI could pass it only if it had acquired adult human intelligence by "experienc[ing] the world as we have."[172] The AI might make the same argument, and contend that the Turing Test was unfair. Would failing the Turing Test be decisive of the question in face of this argument? I suspect not. It would depend on the way that the AI failed the test. French imagines, for example, questions that would detect whether or not the questioned entity had ever baked a cake,[173] but surely a lack of knowledge of experience of cake baking should not disqualify one from the possession of fundamental liberties. Some failures would be relevant, for example failures that indicated that AIs did not possess awareness of themselves as having ends or that they did not understand our words and their own situation.

---

171. Turing, *supra* note 17, at 42.

> The game may perhaps be criticized on the ground that the odds are weighted too heavily against the machine. . . . This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection.

Id.

172. Robert M. French, *Subcognition and the Limits of the Turing Test*, 99 MIND 53, 53-54 (1990).

173. *See id.* at 58.

V. AI REVISITED

My suggestion for an approach to the debate over the possibility of AI can now be restated. Turing, by proposing his test, attempted to operationalize the question whether an AI could think. By borrowing a parlor game as the model for his test, however, Turing failed to provide a hypothetical situation in which outcome of the test had any pragmatic consequence. This failure invites the invention of further hypotheticals, such as Searle's Chinese Room, that add distance between the thought experiment and practical consequences. The result has been that the Turing Test, far from operationalizing the question, has been the occasion for an abstract debate over the nature of "thinking." I propose that we use a different sort of thought experiment: let us modify the Turing Test so that the hypothetical situation focuses our attention on pragmatic consequences. This Essay explored two such thought experiments—the trustee scenario and the constitutional personhood scenario.

These two scenarios raise quite different questions. On the one hand, there is the question whether an artificial intelligence could ever possess the general-purpose competence that we associate with humans. The trustee scenario raises these issues of capacity and responsibility. The focus of the law's inquiry, should the first scenario ever arise, ought to be on whether AIs can function as trustees. "Can an AI do the job?" is the question the law should ask. "Does the AI have an inner mental life?" is simply not a useful question in this context.

On the other hand, there is the question whether an artificial intelligence would have the qualities that give humans moral and legal worth—the kind of value that is protected by social institutions. The constitutional personhood scenario raises these new and different issues. Competence is still relevant, but competence alone is not sufficient to qualify an entity for the rights of constitutional personhood. Intentionality, consciousness, emotion, property rights, humanity—all of these concepts could be relevant to the inquiry.

The difference between these two legal inquiries reveals that there are at least two different issues at stake. When we ask the questions whether a computer running a program could "think," or whether artificial intelligence is possible, the questions are ambiguous. In one sense, an AI would be intelligent if it possessed the sort of all-purpose, independent capacity to function in a role that now requires a competent human adult—trusteeship, for example. In another sense, an AI might not be said to be a "thinking" being, unless it had something like our mental life—unless it possessed consciousness, intentionality, and so

forth. In still a third sense, AIs would not be like us unless they possessed wants, interests, desires, or a good.

Now reconsider the debate over the Chinese Room. Searle's argument that AIs could not possess intentionality seems to be completely irrelevant to the question whether an AI could serve as a trustee. Searle hypothesizes that the person in the Chinese room is perfectly competent at simulating knowledge of Chinese when following the instruction book. Searle might say that the AI could not understand the meaning of the terms of a trust it administered, but he would not question the AI's ability to carry them out. Searle might say that an AI could not understand the meaning of New York Stock Exchange prices, but he does not argue that an AI could not do a better job than a human at investing in the stocks to which those prices relate. If AIs were competent to act as general-purpose trustees, making a wide variety of decisions and responding to novel circumstances, they would be intelligent in a very important sense.

Searle's objection might have some force, however, when it comes to the second scenario—the AI seeking rights of constitutional personhood. In that context, the intentionality objection plays a role similar to the arguments against constitutional personhood based on the premise that an AI would not possess consciousness, intentionality, emotion, or free will. All of these missing-something objections point to the lack of an elusive quality. Flesh and blood can produce intentionality, consciousness, emotion, and free will, but silicon and copper cannot. Of course, Searle did not claim that AIs could not exhibit the behaviors we associate with intentionality (or consciousness and emotion). His point is that these behaviors cannot be evidence of real intentionality.

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. Indeed, if talk about AIs as possessing intentions became a settled part of our way of speaking about AIs, Searle's argument might come to be seen as a misunderstanding of what we mean by "intentionality." If a lawyer brought up Searle's argument in a legal proceeding, some philosophers might say knowingly to each other: "That argument is based on a mistake. Saying that an AI knew where to find a bit of information is a paradigm case of intentionality." Searle can respond by saying that this new way of talking certainly does not reflect what he means by terms such as "intentionality," "knowing," and so forth, and he would be right. But what would be the argument that we should all continue to talk like Searle, long after there will be any reason to do so?

Searle has an answer to this question. Take the example of con-

sciousness rather than intentionality. We might have a reason to deny that AIs possess the kind of consciousness that would count in favor of giving them the rights of constitutional personhood, even though they did a very good imitation of consciousness. Imagine that cognitive science does develop a theory of human consciousness that is confirmed by sufficient evidence, but someone produces an AI that is programmed to produce *only* the recognizable symptoms[174] (and not the real underlying processes) of consciousness. In that case, we would have a good reason *not* to treat the AI as a conscious being. If the illusion of consciousness were a convincing one, we might lapse in our ordinary talk about AIs. Moreover, if we built AIs that seemed conscious, got in the habit of treating them as if they were persons, and then discovered that what they possessed was only a clever simulation of consciousness, we might be quite shocked. Despite the possible shock, our knowledge about how consciousness works would be very relevant and likely decisive for our judgment as to whether an AI had it. Where Searle (or someone who makes a similar argument) goes wrong, I think, is in his insistence that we know enough about consciousness, intentionality, emotion, and free will to rule out the possibility that it can be produced artificially by a computer.

In sum, the two legal scenarios have several implications for the debate over the possibility of artificial intelligence. First, focusing on concrete legal questions forces us to take a pragmatic view of the AI debate; we are forced to consider what hangs on its outcome. Second, the trustee scenario suggests that AIs will need to become very competent indeed before we are tempted to treat them as possessing human-quality intelligence suitable for use as a means to human ends. Third, questions about true intentionality or real consciousness are not relevant to the inquiry in the trustee scenario. Fourth, the constitutional personhood scenario suggests that these questions about mental states will indeed be relevant if we ask whether AIs ought to be treated as ends in themselves. Fifth, the answer to the personhood question is likely to be found two places—in our experience with AI and in our best theories about the underlying mechanisms of the human mind.

Today, we lack both experience with really capable AIs and well-confirmed theories of how the human mind works. Given these gaps, Turing's suggestion that we put aside the question whether AIs can think was a good one. Perhaps we have not put it far enough aside.

174. By symptoms, I mean the surface behaviors—those that could be observed without examining the underlying mechanisms.

VI. PERSONHOOD RECONSIDERED

Finally, I would like to raise some questions about the implications of the AI debate for controversial questions in legal, moral, and political theory. Cognitive science may provide us with a better understanding of our concept of a person. Some of the most intractable questions in jurisprudence, as in ethics and politics, have concerned the borderlines of status—what is a person and why we do give human persons such strong legal protection?[175] Should animals have stronger legal rights? How should we treat criminal defendants with multiple personalities?[176] What should be the legal status of a fetus? Should trees have standing? Many of these questions remain unsettled. Disagreement about their proper answers has persisted and even intensified.

It seems that developments in cognitive science might eventually be brought to bear on some of these questions. For example, it could be argued that personhood is identical with humanity—that possession of the genetic material of homo sapiens is a necessary and sufficient condition for personhood. But cognitive science may give us a very different picture of personhood—a picture that casts doubt on the equivalence between humans and persons. At the other end of the spectrum, AI research might give us insight into the claim that groups have rights that are not reducible to those of individuals.

Thinking about the question whether AIs should ever be made legal persons does shed some light on the difficult questions the law faces about the status of personhood. It is not that we have discovered a theory of personhood that resolves hard questions about the borderlines of status. Rather, thinking about personhood for AIs forces us to acknowledge that we currently lack the resources to develop a fully satisfactory theory of legal or moral personhood.[177] There are reasons for our uneasiness about the hard cases at the borderline of status, and the thought

175. *See* GREENAWALT, *supra* note 109, at 120-43 (discussing the valuation of the life of a fetus).

176. *See* Elyn R. Saks, *Multiple Personality Disorder and Criminal Responsibility*, 25 U.C. DAVIS L. REV. 383 (1992).

177. In this sense, I do not think that considering the philosophical debate about the possibility of AI yields any clear answers for current debates about personhood. *But see* Steven Goldberg, *The Changing Face of Death: Computers, Consciousness, and Nancy Cruzan*, 43 STAN. L. REV. 659, 680 (1991). Goldberg has argued that artificial intelligence research may shape the outcome of at least one legal question about the borderline of status—the definition of death. He begins with the premise that humans have a strong preference for seeing the human species as unique. He then argues that if a social consensus were reached that computers are self-aware, we would then seek another characteristic of humans to distinguish ourselves as unique in the universe. He suggests that this characteristic may be the capacity for social interaction. *Id.* at 680. This leads Goldberg to the conclusion that self-aware computers

experiment in which we have engaged can help us to get a firmer grasp on these reasons.

The first reason for our uneasiness concerns the relationship between our concept of personhood and our concept of humanity. All of the persons we have met have been humans, and the overwhelming majority have been normal humans who give clear behavioral evidence of being conscious, having emotions, understanding meanings, and so forth. Given this coincidence (in the narrow sense), it is not surprising that our concept of person is fuzzy at the edges. For most practical purposes, this fuzziness does not get in our way. We treat humans as persons, and we need not worry about why we do so.

There are, however, occasions on which this strategy fails. Two of the most prominent cases occur at the beginning and the end of human life. Abortion and the cessation of life-sustaining treatment for humans in permanent vegetative states both raise questions about the status of personhood that cannot be answered by a simple comparison with a normal human adult. A third case is that of those higher mammals that seem most likely to have a mental life that is similar to that of humans.

In these cases, we can see the second reason for the persistence of uneasiness about the borderline of personhood. With respect to fetuses, humans in vegetative states, and higher mammals, we lack the sort of evidence we would need to establish a clear-cut case of personhood. Fetuses and humans in permanent vegetative states do not behave as nor-

---

would make it more likely that courts would adopt capacity for social interaction as a definition for death. *Id.* at 681-82.

Goldberg's essay is provocative, but his argument is tenuous. First, Goldberg states but does not argue for the assumption that "any concept of human death depends directly on those qualities thought to make humans unique." *Id.* at 663 (citing ROBERT M. VEATCH, DEATH, DYING AND THE BIOLOGICAL REVOLUTION: OUR LAST QUEST FOR RESPONSIBILITY 29-42 (1976)). Second, Goldberg asserts but does not provide evidence for the proposition that a psychological need for humans to see themselves as unique *caused* the shift from heart-function to brain-function definitions of death. *Id.* at 660-70. Third, Goldberg does not consider other possible chains of causation. Consider two other possibilities. One possibility is that there may have been practical concerns for the cost of sustaining "life" without possibility of recovery. A second possibility is that consciousness may be a condition of personhood, as personhood is understood by the best available moral theory. The development of life-sustaining technology that permits the maintenance of heart function without consciousness for extended periods may have forced the issue of how to define death, which would have only been theoretical before the development of the new technology. Fourth, AIs that are capable of producing consciousness also may be capable of social interaction. Therefore, the same developments in artificial intelligence research that would prompt a move away from a consciousness-based definition of death would also prompt a move away from a social-interaction definition. Fifth, Goldberg's argument seems to imply that the move to a social-interaction definition would be based on a conceptual mistake or some form of wishful thinking. But if Goldberg can see this, why will courts be unable to do so?

mal human adults do, but they are humans.[178] Similarly, we have not been able to communicate with higher mammals in a way that yields clear behavioral evidence of a mental life of human quality, and higher mammals like whales are clearly not humans. In none of these cases is the behavioral evidence sufficient to establish that persons are (or are not) present.

There is a third reason for our persistent doubts about the borderline of personhood. Cognitive science, so far, has not yielded well-confirmed theories of the brain processes that underlie mental states like consciousness, emotion, and so forth. Absent well-confirmed theories of underlying processes, we cannot make confident judgments that the elements of personhood are lacking in particular cases.

Our thought experiment does suggest what sort of evidence might be decisive. If AIs behaved the right way and if cognitive science confirmed that the underlying processes producing these behaviors were relatively similar to the processes of the human mind, we would have very good reason to treat AIs as persons. Moreover, in a future in which we interact with such AIs or with intelligent beings from other planets, we might be forced to refine our concept of person.

The question then becomes what do we do about the hard cases that arise today? Thoughts about the shape of an answer can begin with the nature of justification and argumentation, both moral and legal. Our unreflective intuitions and well-considered moral and legal judgments are rooted in particular cases. These paradigm cases are the stuff of ordinary practical discourse. We make analogies to the familiar cases. We try to bring order to our particular judgments by advancing more general theories. We seek reflective equilibrium between our considered judgments and general theories. Ordinary practical discourse is shallow in the sense that it can be (and usually is) limited to arguments rooted in our common sense and ordinary experience.

What do we do when we must decide a case that goes beyond these shallow waters—the tranquil seas where theories are connected to the ocean floor by familiar examples and strong intuitions? In deep and uncharted waters, we are tempted to navigate by grand theories, grounded on intuitions we pump from the wildest cases we can imagine. This sort of speculation is well and good, if we recognize it for what it is—imaginative theorizing. When it comes to real judges making decisions in real legal cases, we hope for adjudicators that shun deep waters and recoil

178. Fetuses might have some behaviors that are associated with feeling, but they clearly do not engage in behavior that establishes the concurrent presence of consciousness, intentionality, emotion, and free will.

from grand theory. When it comes to our own moral lives, we try our best to stay in shallow waters.[179]

The thought experiments in this Essay have taken us beyond the shallow waters of our intuitions and considered judgments. One way of expressing the result of our journey is this: Our theories of personhood cannot provide an a priori chart for the deep waters at the borderlines of status. An answer to the question whether artificial intelligences should be granted some form of legal personhood cannot be given until our form of life gives the question urgency. But when our daily encounters with artificial intelligence do raise the question of personhood, they may change our perspective about how the question is to be answered.

And so it must be with the hard questions we face today. Debates about the borderlines of status—about abortion, about the termination of medical treatment, and about rights for animals—will not be resolved by deep theories or the intuitions generated by wildly imaginative hypotheticals. Of course, many of us do believe in deep theories; we subscribe to a variety of comprehensive philosophical or religious doctrines. But in a modern, pluralist society, the disagreement about ultimate questions is profound and persistent. Resolution of hard cases in the political and judicial spheres requires the use of public reason. We have no realistic alternative but to seek principled compromise based on our shared heritage of toleration and respect. If there is no common ground on which to build a theory of personhood that resolves a hard case, then judges must fall back on the principle of respect for the rights of those who mutually recognize one another as fellow citizens.

179. *Compare* John Rawls, *The Independence of Moral Theory*, 48 PROC. & ADDRESSES AM. PHIL. ASS'N 5 (1975) (arguing that moral theory should be independent of metaphysics) *with* Robert Stern, *The Relationship Between Moral Theory and Metaphysics*, PROC. ARISTOTLIAN SOC'Y 143 (1992) (arguing that moral theory is dependent upon metaphysics).