Oklahoma Law Review

Volume 72 | Number 1

Symposium: Lawyering in the Age of Artificial Intelligence

2019

Artificial Wisdom? A Potential Limit on AI in Law (and Elsewhere)

Joshua P. Davis

ARTIFICIAL WISDOM? A POTENTIAL LIMIT ON AI IN LAW (AND ELSEWHERE)

JOSHUA P. DAVIS*

Abstract

Artificial intelligence ("AI") may soon perform all tasks that belong to the factual realm more effectively than can human beings. AI may become superior—likely far superior—at describing, predicting, and persuading. Its competitive advantage in these pursuits may well extend to legal and judicial practice. Does that mean human participation in the law will be rendered obsolete? Not necessarily. This Essay suggests three propositions may hold true that would justify an ongoing—perhaps permanent—role for human beings: (1) that moral judgment is necessary for legal and judicial practice; (2) that the first-person perspective (or subjectivity) is necessary for moral judgment; and (3) that AI is incapable of attaining the first-person perspective. After briefly addressing the first two propositions, the Essay focuses on the third. It explores ways in which the best scientific accounts of various phenomena related to the first-person perspective—consciousness, free will, and the unified self—seem incompatible with an internal experience of the first-person perspective, particularly when it comes to decision-making. AI seems to be a creature of science, suggesting it too may be incompatible with the first-person perspective. If so, while we must recognize the staggering potential of artificial intelligence, there is no similar prospect for artificial wisdom. The need for wisdom—understood here as involving moral judgment—preserves a role for human beings in legal decision-making.

Artificial intelligence ("AI") holds extraordinary potential. It has proven superior to human beings in various ways. It plays chess better than we do.[1]

---

* Professor and Director, Center for Law and Ethics, University of San Francisco School of Law. My thinking on these topics has benefited greatly from collaboration with Brad Wendel. I also received valuable comments from Stephen Henderson and Melissa Mortazavi. I am grateful for assistance in research on the topic from one of our excellent research librarians, Suzanne Mawhinney, and from an excellent research assistant, Javkhlan Enkhbayar.

1. *See* MAX TEGMARK, LIFE 3.0: BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE 51 (2017); Dana Mackenzie, *Update: Why This Week's Man-Versus-Machine Go Match Doesn't Matter (and What Does)*, SCI. MAG. (Mar. 15, 2016, 10:00 AM), http://www.sciencemag.org/news/2016/03/update-why-week-s-man-versus-machine-go-match-doesn-t-matter-and-what-does (discussing the World Chess Champion's loss to a computer in 1997).

It plays Go better than we do.[2] It will soon drive cars better than we do.[3] In the not too distant future, it may well be able to program computers better than we can.[4] When that occurs, many experts predict a rapid acceleration.[5] AI will create enhanced AI, which will create even more enhanced AI, and a recursive loop will follow. Through a process that will progress at an exponential rate, the result will be artificial superintelligence ("ASI").[6] Neither the speed nor the limit of this process will be linear. ASI in rapid succession will exceed the intelligence of any human being, then the collective intelligence of all human beings, and then perhaps any level of intelligence we can imagine. Enter the technological singularity. As we contemplate this possibility—inevitability?—we must shift from asking what AI can do to asking what, if anything, it cannot do.

This Essay proceeds from the assumption that ASI will master the world of fact. It will build machines that can sense anything any current life form can, and perhaps much that eludes the perceptions of life as we know it today. ASI will also solve any scientific problem that we have the potential to solve. Human beings will be forced to cede the realms of description and prediction. Any enduring human participation in those efforts will be a pastime. Some people now drive restored cars from the 1950s. Others collect manual typewriters. But those hobbies do not contribute meaningfully to

---

2. *World's Best Go Player Flummoxed by Google's 'Godlike' AlphaGo AI*, GUARDIAN (May 23, 2017, 6:29 AM), https://www.theguardian.com/technology/2017/may/23/alphago-google-ai-beats-ke-jie-china-go. The best Go player in the world, Ke Jie, thought he would never lose to a computer. *Id.* His response: "I feel like his game is more and more like the 'Go god.' Really, it is brilliant." *Id.*

3. *See, e.g.*, JERRY KAPLAN, ARTIFICIAL INTELLIGENCE: WHAT EVERYONE NEEDS TO KNOW 41-42 (2016).

4. As John Searle once commented,

> [A]s soon as someone says that there is a certain sort of task that computers cannot do, the temptation is very strong to design a program that performs precisely that task. And this has often happened. When it happens, the critics of artificial intelligence usually say that the task was not all that important anyway and the computer successes do not really count. The defenders of artificial intelligence feel, with some justice, that the goal posts are being constantly moved.

JOHN SEARLE, MIND: A BRIEF INTRODUCTION 63 (2004) [hereinafter SEARLE, MIND: A BRIEF INTRODUCTION].

5. *See, e.g.*, KAPLAN, *supra* note 3, at 138-43; Irving John Good, *Speculations Concerning the First Ultraintelligent Machine*, *in* ADVANCES IN COMPUTERS 31, 78 (Franz L. Alt & Morris Rubinoff eds., 1965).

6. *See, e.g.*, NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES (2014); ROMAN V. YAMPOLSKIY, ARTIFICIAL SUPERINTELLIGENCE: A FUTURISTIC APPROACH (2015).

technological advancement. All human participation in science will be a hobby. If any of us understands the mechanics of what ASI achieves, it will be because ASI uses its genius to provide dumbed-down explanations so that the most brilliant among us can glimpse in simplistic form what none of us can comprehend. Even they will be privileged spectators, not meaningful participants.

The resulting risk is that the whole working world will follow taxi drivers and truck drivers into the dustbin of obsolescence.[7] ASI will diagnose illnesses, prescribe medications, and perform surgeries far better than doctors. It will design and assemble buildings far better than architects and construction workers. And it will be able to describe the law, predict how people will interpret and apply the law, and frame persuasive legal arguments far better than judges and lawyers. Does that mean ASI will displace all legal practitioners? Not necessarily.

Facts may not exhaust what exists in the world. There is also the realm of value. While AI is capable of helping us pursue our ends, it is not clear that AI can choose ultimate ends for us. The axes of fact and value may be orthogonal—extending at a right angle from one another. As Hume long ago suggested, it may not be possible to derive an "ought" from an "is."[8] If not, then AI—and ASI—may not be able to make any progress along the axis of value, no matter its extension along the axis of fact. Increasing zero by any factor or exponent still leaves zero.

This Essay explores the possibility that ASI will not be able to conquer the realm of value and that its inability to do so will circumscribe the role it can play in law. The Essay does not offer a fully developed argument but rather the beginnings of one. It starts with intuitions about the limits of AI (and ASI). Two characteristics—likely related—mark AI. First, it appears to operate purely in the realm of the scientific world. The knowledge it captures is most consistent with a naturalized materialism—with a popular understanding of science. Second, at least as far as we know, AI inhabits a third-person perspective. We have no solid basis thus far to believe AI—or even ASI—can achieve consciousness, exercise free will, experience a unified self, or otherwise embody subjectivity, however that might best be understood. This third-person, scientific perspective—so the intuition runs—

---

7. Andrew G. Simpson, *4 Million Driving Jobs at Risk from Autonomous Vehicles: Report*, INS. J. (Mar. 27, 2017), https://www.insurancejournal.com/news/national/2017/03/27/445638.htm.

8. *See, e.g.*, SCOTT J. SHAPIRO, LEGALITY 47 (2011) (discussing this position by Hume). For a general discussion of the distinction between the realms of fact and value, see RONALD DWORKIN, RELIGION WITHOUT GOD (2013).

may not be capable of selecting ultimate ends[9] as opposed to identifying and pursuing means.

This limitation holds great importance if first-person decision-making—or evaluation—is necessary to make moral and other value judgments. There is reason to believe that the scientific perspective cannot fully capture first-person decision-making.[10] One way to assess this potential limitation is by reviewing how the scientific view has attempted to make sense of various phenomena relevant to the first-person perspective. These phenomena include consciousness, free will, and the unified self. What we may find is that the scientific worldview cannot offer an account of these phenomena capable of explaining and guiding first-person decision-making. Science and first-person decision-making, it seems, may be somewhat incompatible.

That incompatibility would be significant. A common—and particularly compelling—justification for the scientific worldview is pragmatism.[11] Science works. It makes planes that fly, lights that illuminate, and telephones that communicate. Science's practical utility provides reason to treat naturalized materialism as a form—perhaps even the only form—of knowledge. But if value judgments are necessary to direct our scientific and other efforts, if first-person decision-making is necessary for value judgments, and if science cannot fully capture first-person decision-making, then science's utility has constraints. Pragmatism may justify recognizing not only a scientific realm of knowledge but also an outer boundary to that realm. Treating scientific knowledge as the only form of knowledge may leave us without the resources we need to make the decisions we need to make.

This sketch of a general philosophical claim finds a special application in law. To the extent that legal practice requires moral (and other value) judgments, a purely scientific account may not exhaust legal decision-making. Science may not be capable of creating machines that have the first-person experiences necessary to exercise key judgments in the legal process. And there is reason to believe that legal practice does require moral and other

9. I say ultimate ends—as opposed to ends—because given ultimate ends, AI may be able to identify intermediate ends that can help to achieve the ultimate ends. *See, e.g.*, TEGMARK, *supra* note 1, at 264 (discussing the relationship of subgoals and ultimate goals); BOSTROM, *supra* note 5, at 132 (discussing the instrumental convergence thesis). This point retains the distinction between fact and value. AI could identify intermediate ends relying solely on ultimate ends provided to it and instrumental reasoning of a factual nature.

10. For an analysis that seems to hold this implication, see THOMAS NAGEL, MIND AND COSMOS: WHY THE MATERIALIST NEO-DARWINIAN CONCEPTION OF NATURE IS ALMOST CERTAINLY FALSE (2012).

11. *See, e.g.*, BRIAN LEITER, NATURALIZING JURISPRUDENCE: ESSAYS ON AMERICAN LEGAL REALISM AND NATURALISM IN PHILOSOPHY 48-50 (2007).

values judgments. According to natural law, saying what the law is (at least sometimes) requires making moral judgments about what the law should be. To be sure, some legal positivists deny that morality plays that role in the law. But in defending that position, even legal positivists generally acknowledge that moral judgments are often necessary in *applying* the law.[12]

These final points take us full circle. They suggest several conclusions: judicial and other legal practitioners may need to be able to make moral and other value judgments; subjectivity may be necessary to moral judgments, restricting AI to the realm of science and precluding it from the realm of value; and science may be incapable of fully capturing the kind of first-person decision-making that is necessary for moral and other value judgments. As far as we can tell, then, human beings may be uniquely capable of a first-person perspective that is necessary to certain forms of judicial and other legal practice. According to this line of reasoning, human beings will have a role to play in the practice of law at least commensurate with the role of moral and other value judgments.

Ultimately, an argument about the limits of AI in law along the above lines would need to support three claims. First, moral judgments (and other value judgments) are necessary for legal and judicial practice. Second, the subjective perspective is necessary for such judgments. Third, computers (and the material world in which they function) cannot achieve subjective experience. This Essay does not make a full version of any of those arguments. Its aim is more modest. Part I explores why it is likely that moral judgments (and other value judgments) may be necessary for legal and judicial practice. Part II explains how the subjective perspective may be necessary for moral judgments and how that may be consistent with morality nonetheless remaining objective in an important sense. Part III then provides evidence that computers may be incapable of subjective experience.

I. Moral Judgments as Necessary in Legal and Judicial Practice

A. Moral Judgment in the Law

Moral judgments may require the first-person perspective—while possibly nonetheless being objectively right or wrong—and AI may be incapable of achieving the first-person perspective and therefore of making moral judgments. If both propositions are true, then the scope of what AI can do in the law would seem to depend on a longstanding and central dispute in

12. *See infra* Part I.

jurisprudence: the role of morality in law.[13] But it turns out that is only partially accurate. It is true that natural lawyers, according to one formulation, contend that the content of law depends ultimately in part on moral judgments.[14] It is also true that legal positivists, according to the same formulation, deny that the content of the law depends ultimately on moral judgments at all.[15] But many of the most prominent jurisprudents—natural lawyers and positivists alike—acknowledge that morality plays some role in legal and judicial practice.

To see why this is so, it is worthwhile to review briefly the views of different schools of thought about the central point of contention in jurisprudence for the past fifty years or more: the relationship between law and morality. A relatively straightforward account of the dispute includes three positions: that morality plays a necessary role, a contingent role, or no role at all. The position that morality plays a necessary role in legal interpretation is often called natural law or anti-positivism. Among its most famous recent exponents were Ronald Dworkin and Lon Fuller.[16] Although their theories vary a great deal in the particulars and emphasis, they both subscribed to the view that moral judgment is necessary in saying what the law is. If computers are unable to engage in moral reasoning, and if natural law provides the best understanding of the nature of law, then computers can play only a limited role in legal interpretation.

To be sure, that does not mean that computers can play no role at all. Take, for example, the theoretical framework that Ronald Dworkin developed. He viewed legal interpretation as entailing two forms of nested judgments: fit

---

13. This debate is often labeled the Hart-Dworkin debate, and it has been the main one in jurisprudence. *See, e.g.*, Scott Hershovitz, *The End of Jurisprudence*, 124 YALE L.J. 1160, 1162 (2015) ("For more than forty years, jurisprudence has been dominated by the Hart-Dworkin debate."); Scott J. Shapiro, *The 'Hart-Dworkin' Debate: A Short Guide for the Perplexed*, *in* RONALD DWORKIN 22 (Arthur Ripstein ed., 2007) ("For the past four decades, Anglo-American legal philosophy has been preoccupied—some might say obsessed—with something called the 'Hart-Dworkin' debate."). Hart and Dworkin's disagreement was germinal of ongoing disputes about legal positivism, with Hart likely the most influential positivist of the past half century and Dworkin the most influential non-positivist.

14. *See* Joshua P. Davis, *Legality, Morality, Duality*, 2014 UTAH L. REV. 55, 61-62 [hereinafter Davis, *Legality, Morality, Duality*]; *see also* Amy Salyzyn, *Positivist Legal Ethics Theory and the Law Governing Lawyers: A Few Puzzles Worth Solving*, 42 HOFSTRA L. REV. 1063, 1063 (2014) (describing positivism as concept that, "broadly speaking . . . informs a particular view of the lawyer as governed in her actions by the legal entitlements at issue, as opposed to, for example, considerations of morality or justice writ at large").

15. *See* Davis, *Legality, Morality, Duality*, *supra* note 14, at 61-62.

16. *See, e.g.*, RONALD DWORKIN, JUSTICE FOR HEDGEHOGS (2011) [hereinafter DWORKIN, JUSTICE FOR HEDGEHOGS]; LON L. FULLER, THE MORALITY OF LAW (1969).

and justification.[17] Fit is a purely descriptive (or positive) assessment.[18] It asks how well a legal proposition fits the existing authoritative sources of law. Justification, in contrast, requires a prescriptive (or normative) assessment.[19] The relevant inquiry is how normatively attractive various competing legal propositions are that potentially fit the relevant authoritative sources of law. So even under Dworkin's anti-positivist jurisprudence, computers might play a significant role. They may someday—perhaps someday soon—be better than human beings at assessing fit. But if they cannot make moral judgments, they will not be able to assess justification.

Moreover, as noted above, assessments about fit and justification are "nested." That is to say, the two kinds of judgments are not discrete. So, for example, a judge needs to make relatively general or abstract assessments of both fit and justification in deciding how she should go about interpreting the law. Both kinds of judgment are relevant to determining whether, for example, in interpreting a statute, a judge should consider its plain text, its legislative history, the overall structure of the statutory system, or which interpretation would make the legal system as just as possible. Only after making those general assessments of both fit and justification will she be in a position to make the more specific (or concrete) assessments about, for example, the text of the statute and how just the potential interpretations of that text would be.[20] So one must be careful not to exaggerate the ease of disentangling the judgments about fit and justification. Still, even in Dworkin's natural-law framework, a robo-judge unable to make moral judgments at least in theory—and, likely, to some extent in practice—might be asked to make useful judgments about fit.

A second jurisprudential position holds that morality plays a contingent—not a necessary—role in saying what the law is. That school of jurisprudence is known as inclusive legal positivism. It can be described as holding that the content of law is dependent ultimately only on social facts but that those social facts may permit or require legal interpreters to make moral judgments, at least in some circumstances. The most famous modern inclusive legal positivist was H.L.A. Hart,[21] and the school includes other major figures,

17. *See, e.g.*, Joshua P. Davis, *Cardozo's Judicial Craft and What Cases Come to Mean*, 68 N.Y.U. L. REV. 777, 809 (1993).

18. *See* RONALD DWORKIN, LAW'S EMPIRE 242-50 (1986).

19. *See id.* at 243-44.

20. *See, e.g.*, Davis, *Legality, Morality, Duality*, *supra* note 14, at 94-95.

21. *See, e.g.*, H.L.A. HART, THE CONCEPT OF LAW (1961).

such as Jules Coleman.[22] According to inclusive legal positivism, it may be a matter of social fact, for example, whether interpreting the Eighth Amendment's ban on cruel and unusual punishment entails only a factual inquiry—perhaps into the practices at the time the amendment was adopted—or whether it entails also a moral inquiry—perhaps whether a practice today is in fact cruel. That social fact determines whether judges must make moral judgments in interpreting the law. So according to inclusive legal positivism, even if computers cannot make moral judgments, it may be a contingent matter whether they will be able to interpret the law in any given instance.

The third major jurisprudential position—exclusive legal positivism—holds that moral facts do not play a role in determining the content of the law. The content of law, according to this position, is purely a matter of social fact. Major modern figures subscribing to this view include Joseph Raz,[23] Scott Shapiro,[24] and Brian Leiter.[25] Exclusive legal positivism appears most compatible with computers serving as judges or lawyers, at least if they cannot make moral judgments.[26] But that appearance may be deceiving.

The reason is that there is more to legal or judicial practice than legal interpretation. Engaging in legal practice—whether as a judge or a lawyer—may involve moral judgments, even if saying what the law is does not. The major exclusive legal positivists have acknowledged this point. Consider Joseph Raz. His jurisprudential position relies on a distinctive understanding of the nature of authority. He claims the point of authority is to resolve moral disputes.[27] So, he contends, it is implicit in the nature of legal authority that authoritative sources of law—*e.g.*, statutes—preclude legal interpreters from addressing the same underlying moral issues that the authoritative sources

22. *See, e.g.*, Jules L. Coleman, *The Architecture of Jurisprudence*, 121 YALE L.J. 2 (2011).

23. *See, e.g.*, JOSEPH RAZ, BETWEEN AUTHORITY AND INTERPRETATION: ON THE THEORY OF LAW AND PRACTICAL REASON (2009).

24. *See, e.g.*, SHAPIRO, *supra* note 8, at 47.

25. *See, e.g.*, LEITER, *supra* note 11.

26. One other jurisprudential school of thought worth mentioning is normative (or ethical) legal positivism. Joshua P. Davis, *Legal Ethics, Legal Dualism, and Fidelity to Law*, 2016 J. PROF. LAW. 1, 8 [hereinafter Davis, *Legal Ethics*]. It holds that the law may reflect moral (or other normative) judgments at its foundation but only at its foundation. *Id.* Jeremey Waldron has described this theory. *See id.* at 8 n.37 (citing Jeremey Waldron, *Normative (or Ethical) Positivism, in* HART'S POSTSCRIPT: ESSAYS ON THE POSTSCRIPT TO THE *CONCEPT OF LAW* 411 (Jules Coleman ed., 2001)). Justice Scalia arguably fit into this category. Id. at 12-13. Note that, according to the taxonomy in the text, that would make Justice Scalia a natural lawyer.

27. RAZ, *supra* note 23, at 4-7.

were meant to resolve.[28] But he does not deny that morality can play a role in legal practice.[29] His position instead is that even if morality informs legal practice, it is not part of the law. He thus distinguishes between "the law of the land"[30]—the law in a particular jurisdiction—and "standards that the courts have to apply"—which are not part of the law of the land. He illustrates his point by drawing an analogy to a British court applying Polish law under British conflict-of-law rules.[31] Foreign law, Raz reasons, may play a role in legal practice. A judge may apply foreign law as a result of domestic choice-of-law rules, just as an attorney may do so. But that does not make foreign law a part of domestic law. So it is with morality, Raz reasons.[32] Whatever the merits of this position—or of Raz's theory of authority[33]—the point that is relevant for present purposes is that Raz's exclusive legal positivism does not preclude judges or lawyers from making moral judgments in applying the law, as opposed to interpreting it.

The same is true for Scott Shapiro's jurisprudential theory, which has important parallels to Raz's. According to Shapiro, law is a kind of plan (or a plan-like norm).[34] Law as a plan provides a way to contend with various difficulties that arise in complex societies, like ours. We face numerous serious moral problems without obviously correct solutions, and yet we need to resolve them—at least provisionally—so we can coordinate our behavior and organize our lives. Shapiro labels this situation "the circumstances of legality."[35] He claims that law as a plan (or plan-like norm) enables coordination and at least temporary resolution of various moral disputes. Exercising moral judgment in saying what the law is, he further reasons, would defeat the purpose of law, given that its aim is to resolve moral disagreements (again, at least temporarily) so that members of society can plan their actions and go about their business. Much like Raz's authority theory, Shapiro's planning theory purports to preclude a role for morality in saying what the law is, but it leaves room for moral judgments in legal practice, including in applying the law. Shapiro acknowledges, for example, that a legal practitioner applying the legal standard of unreasonableness or

28. *Id.* at 190-93.
29. *Id.* at 195-98.
30. *See id.* at 198-99.
31. *Id.* at 199.
32. *Id.* at 201.
33. For a critique of Raz's argument for exclusive legal positivism based on the notion of authority, see RONALD DWORKIN, JUSTICE IN ROBES 198-211 (2006).
34. SHAPIRO, *supra* note 8, at 225.
35. *Id.* at 170-73, 213-14.

unconscionability may have to make a moral judgment.[36] He also invokes Raz's distinction between the laws of a jurisdiction and the "rules that judges are under an obligation to apply."[37] The latter category, he claims, is bigger, including the rules of English grammar and the laws of foreign jurisdictions.[38]

Legal positivists—exclusive and inclusive alike—also acknowledge that judges may exercise moral judgment when they make new law. According to some legal positivists, when the law is indeterminate—jurisprudents dispute how often that occurs—judges may have to make moral judgments in formulating the governing legal standard. Shapiro, for example, acknowledges that judges sometimes make new law and that when they do, they may rely on moral judgments.[39] The plan that is the law has not contemplated a situation that has arisen. The compromise among competing moral values and interests that society has adopted does not provide determinate guidance. Recourse to underlying moral values may be necessary to extend the law.[40]

Hart too recognized that the law sometimes is indeterminate. He thus distinguished between cases located in the "core" of the law and those located in its "penumbra."[41] Core cases lie within the settled meaning of the law while penumbral cases require recourse to social purposes to reach a determinate result.[42] To be sure, Hart denied that the social purposes that guide decision-making in the penumbra of the law must be moral.[43] Not only

---

36. *Id.* at 276.

37. *Id.* at 272.

38. *Id.*

39. *Id.* at 272, 274-76. For a discussion of the role that Shapiro admits morality can play in adjudication, and for the challenges that admission may raise for his theory, see Davis, *Legality, Morality, Duality*, *supra* note 14, at 76-80.

Hart's analysis is somewhat more complicated. He famously acknowledged that legal interpreters must make judgments about the ends or purposes of law when cases fall in its "penumbra." *See* H.L.A. Hart, *Positivism and the Separation of Law and Morals*, 71 HARV. L. REV. 593, 607 (1958). Whether those judgments are necessarily moral in character, however, he disputed. *Id.* at 614. Note that this point can tie into the discussion of other value judgments that also may be beyond the scope of AI.

40. SHAPIRO, *supra* note 8, at 274-76; *id.* at 272 ("In other words, the fact that American judges are under an obligation to apply nonpedigreed norms does not imply that they are compelled to apply preexisting law; rather, they are merely under an obligation to reach outside the law and apply the norms of morality instead.").

41. Hart, *supra* note 39, at 607. Hart claimed that Austin recognized this issue as well. *Id.* at 608-09.

42. *Id.* at 612, 614.

43. *Id.* at 612-13.

moral aims can provide guidance when the law is indeterminate, according to Hart, but so can "the most evil aims."[44] Not all jurisprudents agree with this point.[45] Assuming Hart was right, however, the conclusion does not necessarily follow that computers can—or will be able to—make relevant non-moral value judgments. Computers may have as great difficulty assessing evil aims as good ones. Indeed, evil values can be understood as a mirror image of good ones—or, if one prefers, as a photographic negative. In that case, the same difficulties that impede robojudges or robolawyers in pursuing moral aims may confound them in pursuing immoral ones. And the same may be true for values that run orthogonal to good and evil—amoral aims. In any case, the relevant aims of the law presumably are at least at times moral, and when they are, computers would need to be able to make moral judgments to engage in effective legal practice.

The point is that proponents of all of the major jurisprudential schools recognize that morality can play some role in judicial and legal practice. They may disagree on whether and when moral judgments are necessary to say what the law is, but they agree that lawyers and judges at times must make moral judgments in going about their business. So if AI is not capable of making moral judgments, it cannot fully perform the role that people play in the law.[46]

---

44. *Id.* at 613; *see also id.* at 629.

45. *See, e.g.*, Lon L. Fuller, *Positivism and Fidelity to Law—A Reply to Professor Hart*, 71 HARV. L. REV. 630 (1958).

46. Resolving jurisprudential disagreements would likely still matter for defining the scope that AI can play in law. If, for example, AI cannot make moral judgments but it can make all other necessary judgments, then exclusive legal positivists might conclude that AI can say what the law is whereas natural lawyers might conclude that AI cannot do so. My own view—explored elsewhere—is that legal positivism provides the best understanding of the nature of law when an interpreter seeks merely to describe the law or to predict how others will interpret it, but natural lawyers provide the best understanding of the nature of the law when interpreters look to the law as a source of moral guidance. *See, e.g.*, Davis, *Legal Ethics*, *supra* note 26; Davis, *Legality, Morality, Duality*, *supra* note 14. That jurisprudential theory can provide a useful account of the potential and limits for AI in law: as a powerful tool for describing and predicting from the third-person perspective but as a limited tool for making legal judgments from the first-person perspective.

A potential response to my argument could be to challenge the strong distinction I assume exists between moral and non-moral claims. Along these lines, consider where Jules Coleman ends his analysis of the architecture of jurisprudence. He suggests the importance for legal positivism of challenging Hume's Law—that one cannot derive an "ought" from an "is." Coleman, *supra* note 22, at 77-78. That is a crucial question from my perspective, both for jurisprudence in general and for the role of AI in particular. Can AI derive an "ought" from an "is"? If not, moral (and other value) judgments may place an outer limit on what computers can do. If so, perhaps there is no such limit.

B. Reciprocity and Moral Judgment

There is another reason to conclude that those who make legal judgments must be capable of making moral judgments. It arises from what one might call a principle of reciprocity. The logic is that those who impose legal judgments on others must themselves be subject to the law. The law, in turn, applies only to conscious actors capable of moral agency. We do not, for example, allow criminal prosecutions of animals or inanimate objects.[47]

Scholars writing about AI and ethics have taken various approaches to the principle of reciprocity. Bradley Wendel, for example, relies on ethical theory. He writes of law as a way of imposing obligations, one that by its nature requires mutuality. As he explains, "The law is a means for giving the types of reasons that human moral agents owe to one another, in response to others' demands for accountability."[48] According to Wendel, the process of reason giving necessarily occurs in the second person.[49] It is relational. Thus, those who make the relevant legal judgments—judges, but also at times lawyers—must be moral agents themselves.[50]

Kiel Brennan-Marquez and Stephen Henderson make an argument sounding in democratic theory to arrive at a similar conclusion. They contend that it is essential in a democracy that the entity passing judgment under the law could be subject to the law and vice versa.[51] In that sense, the law is "self-imposed."[52] A judge—or, again, potentially a lawyer—must be able to say, "There but for the grace of God go I" (or the secular equivalent). They characterize their position as requiring "role-reversible judgment."[53]

---

47. We similarly do not allow criminal prosecution of people with such diminished capacity that they have no ability to tell right from wrong, although subtleties arise about whether a defendant must be able to assess legality as opposed to morality. *See, e.g.,* State v. Winder, 979 A.2d 312, 318 (N.J. 2009). But note the complexities for applying the law to people with diminished capacity. Note also that the very different demographics for areas of the law—e.g., judges versus criminal defendants—cause difficulties, requiring an idealized understanding of reciprocity.

48. W. Bradley Wendel, *The Promise and Limitations of Artificial Intelligence in the Practice of Law*, 72 OKLA. L. REV. 21, 42 (2019); *see also id.* at 26-27.

49. *Id.* at 26-27, 42.

50. *Id.*

51. Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137, 149-56 (2019).

52. *Id.* at 153.

53. *Id.* at 149-52.

Note that both of these lines of analysis assume that AI cannot properly be a moral agent.[54] At present, that seems like a reasonable assumption. Computers do not appear to have moral agency or to warrant the equal concern and respect that human beings do. But could they in the future? The following Parts suggest reason to doubt that they will. Toward that end, Part II explains why the first-person perspective is likely necessary for moral judgment, and Part III provides reason to believe AI will not be able to acquire a first-person perspective.

II. The First-Person Perspective as Necessary for Moral Judgment

> *It is the star to every wand'ring bark,*
> *Whose worth's unknown, although his height be taken*.[55]
>
> Russian gentleman: *So who is to say what is moral?*
>
> Sonja: *Morality is subjective.*
>
> Russian gentleman: *Subjectivity is objective.*
>
> Sonja: *Moral notions imply attributes to substances which exist only in relational duality.*
>
> Russian gentleman: *Not as an essential extension of ontological existence.*
>
> Sonja: *Can we not talk about sex so much?*[56]

A. Reasons to Believe a Subjective Perspective Is Necessary for Morality

Subjectivity may be necessary for morality because morality involves evaluation and choice in the realm of value. There is of course a great deal of controversy about what moral judgments are and how one should go about making them. What is at least somewhat less controversial is that someone—with a distinct point of view—needs to make moral judgments.[57] This is most

54. For an argument that focuses even more on process—based on an understanding of litigation as integral to democracy—see ALEXANDRA LAHAV, IN PRAISE OF LITIGATION (2017). Lahav's argument provides another potential reason to limit the role of AI in litigation, reserving a role for citizens and other participants in a healthy democratic process. That said, if AI attains the first-person perspective, perhaps it should be able to participate in democratic processes and, to that extent, in litigation.

55. WILLIAM SHAKESPEARE, SONNET 116.

56. LOVE AND DEATH (Jack Rollins & Charles H. Joffe Productions 1975).

57. *See, e.g.*, SEARLE, MIND: A BRIEF INTRODUCTION, *supra* note 4, at 110 ("One of the weird features of recent intellectual life was the idea that consciousness—in the literal sense of qualitative, subjective states and processes—was not important, that somehow it did not

obvious if moral claims are really just statements of desires or preferences.[58] Disembodied desires and preferences do not exist in the world. Someone must have them.

Even if morality involves more than just desires or preferences—even if moral claims can in some sense be true or false—subjectivity is likely essential in making moral judgments. One way to get at this point is through the fact-value distinction. In many—but not all—accounts of morality, values do not exist in the physical world.[59] They are not discoverable through the empirical sciences—although empirical facts are relevant to moral judgments.[60] On these accounts of morality, a point of view is necessary to identify moral principles or to assess the morality of particular actions.

We find some confirmation for this view in the current state of AI. To date, it cannot select its own ultimate ends, moral or otherwise. Human beings must define the goals AI will pursue. When it comes to "machine morality"—to the possibility of AI pursuing moral ends—that leaves two main options: a top-down approach[61] and a bottom-up approach.[62] First, in the top-down approach, human beings can identify general moral principles for AI. Once those are in place, they can constrain or direct what AI will attempt to achieve. But AI cannot supply its own guiding principles or ultimate aims. It operates in the realm of means, not ends.

---

matter. One reason this is so preposterous is that consciousness is itself the condition of anything having importance. Only to a conscious being can there be any such thing as importance."); *see also* JOHN SEARLE, MIND, LANGUAGE, AND SOCIETY 83 (1998) [hereinafter SEARLE, MIND, LANGUAGE, AND SOCIETY] ("Any attempt to describe consciousness, any attempt to show how consciousness fits into the world at large, always seems to me inadequate. What we are leaving out is that consciousness is not just an import feature of reality. There is a sense in which it is *the* most important feature of reality because all other things have value, importance, merit, or worth only in relation to consciousness. If we value life, justice, beauty, survival, reproduction, it is only as conscious beings that we value them.").

58. This view is often labeled moral nihilism. *See, e.g.*, RUSS SHAFER-LANDAU, THE FUNDAMENTALS OF ETHICS 292 (3d ed. 2015).

59. For a suggestion that morality may be objective but not reducible to at least current understandings of scientific naturalism, see NAGEL, *supra* note 10, at 97-126 (chapter 5, "Value").

60. Note that a naturalized moral objectivity provides an exception to this view. For an analysis along those lines, see PHILIP KITCHER, THE ETHICAL PROJECT (2011). This Essay assumes that values cannot be reduced to the hard sciences. But an argument in support of that position is beyond its scope.

61. WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG 83-97 (2009).

62. *Id.* at 99-115. Wallach and Allen also discuss a hybrid of the two. *Id.* at 117-24.

In the second, bottom-up approach, AI is provided data about various situations and the moral (or otherwise desirable) actions to be taken in them. AI can then detect patterns of moral decision-making and use them for guidance. Note in this approach, too, AI needs input from human beings. Someone with a point of view must determine what the moral actions are in particular situations. Without those judgments, AI would not have the data necessary to detect patterns.

To be sure, there are other forms of moral judgment than either a pure top-down or bottom-up approach. One of the most widely recognized is the so-called "reflective equilibrium," famously discussed by John Rawls, among others.[63] A reflective equilibrium can be understood as involving working back and forth between general moral principles and intuitions about morality in particular contexts.[64] Each informs—and, ideally, corrects—the other. For AI to pursue a reflective equilibrium, human beings would seem to have to provide it at least three forms of guidance: a set of preliminary general principles, a set of intuitions about the right outcomes in particular cases, and a set of rules for reconciling conflicts or tensions between the two.

It is possible that once AI has the necessary inputs from human beings, it could then make moral judgments independently. But there is reason to doubt that would work in practice. One difficulty arises from change. New circumstances may require new moral judgments. Of course, if the relevant differences are reflected in the old data—if the salient distinctions have been judged in the past—AI should be able to contend with them. But sometimes new facts require novel moral judgments to place them within existing moral frameworks. So as circumstances change over time, new human judgments may be necessary for AI to adapt.

Moral values may also change. The reasons are not clear. Perhaps there is moral progress. We may come to be more enlightened as the years pass. Or perhaps values simply alter over time—not necessarily for better or for worse. Regardless, morality does not seem to remain constant. As morality changes, new human input will be necessary for AI to remain accurate in its moral judgments. Human beings will need to alter or add to the moral principles on which AI operates, the concrete moral judgments AI treats as correct, or both.

Another, related difficulty involves error. The moral principles human beings articulate, and the concrete moral judgments they make, are likely to

63. *See, e.g.*, JOHN RAWLS, A THEORY OF JUSTICE (1971).

64. Rawls himself avoided use of the term intuition in this context, and whether doing so is proper remains controversial. *See Reflection Equilibrium*, STAN. ENCYCLOPEDIA OF PHIL. (Oct. 14, 2016), https://plato.stanford.edu/entries/reflective-equilibrium/.

be imperfect. To the extent they are tainted—by ignorance, confusion, self-interest, simple mistakes, or whatever else—they will lead AI astray. As the saying in statistics goes: garbage in, garbage out. To the extent that AI derives its principles either directly from human instruction or from patterns of human moral decisions, both may contain inaccuracies. Human intervention may be necessary to correct the mistakes embedded in AI ethics by human fallibility.

All of these issues will be expected to arise as long as AI engages in moral reasoning at a level of remove: acting on human perceptions of morality rather than on its own perceptions. So if it is correct that the subjective perspective is necessary to make moral judgments, AI would likely need to achieve subjectivity to displace the human role in moral reasoning.[65]

To the extent AI cannot make moral judgments directly, the discussion in Part I suggests two limits for AI in law. First, human beings may need to play some ongoing role in law because only they can assess morality directly. They must provide the moral reasoning or intuitions to feed AI's legal analysis, at least when morality informs judicial or other legal decision-making. Second, human beings may need to play a role to the extent that actors in the legal system have to be capable of moral agency. An example is the principle of reciprocity discussed above—championed in different forms by Wendel and by Brennan-Marquez and Henderson. If AI cannot make direct moral judgments, it would seem incapable of the kind of moral agency that the reciprocity principle requires.[66] Before addressing reasons AI might not be able to achieve subjectivity—and therefore might not be able to make direct moral judgments—it is important to clarify different ways to understand subjectivity and objectivity in morality.

B. How Morality Might Be Both Subjective and Objective

The notion that a subjective perspective is necessary for moral judgment can be misleading. One might take it to mean that morality is subjective. But that is not necessarily true. The reason is that the term subjectivity can have

65. There may also be an issue with coherence. It may be important that moral judgments are coherent. *See, e.g.*, AARON ZIMMERMAN, MORAL EPISTEMOLOGY 11 (2010). Synthesizing moral principles and judgments from people with different perspectives may not generate a single, coherent moral worldview. For related analyses of the limited potential of AI to make moral judgments, see Joshua P. Davis, *Law Without Mind: AI, Ethics, and Jurisprudence*, 55 CAL. W. L. REV. 165, 186-95 (2018).

66. Even if AI could make direct moral judgments, it still might lack moral agency. Other requirements of moral agency might include consciousness or free will. As discussed in Part III, there is reason to doubt AI will be able to achieve either of those as well.

multiple meanings, at least two of which are relevant for present purposes.[67] The first is that a subjective perspective is necessary to make moral judgments. The second is that there is no objectively correct view about whether a moral claim is right or wrong. It is possible that morality may be subjective in the first sense—subjective experience may be necessary to make moral judgments—but not in the second sense—there may be right and wrong answers to (at least some) moral questions.[68]

Of course, to say that something is conceptually possible is not to say that it is true. This Section makes a claim about the former, not the latter. The issues this Part briefly explores are highly controversial. Philosophers—and others—disagree about the ways in which subjective experience is necessary to make moral judgments and also about whether moral judgments can be objectively true. This Part does not attempt to resolve either controversy. Instead, it seeks to describe a view that could combine subjectivity in one regard with objectivity in the other regard.

It is also true that some philosophers—and others—relate the two issues above to each other. They believe that whether moral judgments can be objectively true depends in part on whether they can be tested from what one might call a third-person perspective. Objective truth, according to a version of this view, might require that there be some sort of moral entity that science can detect and measure. From this vantage, it may be that if morality is subjective in the sense that moral judgment requires a subjective perspective, it follows that morality is also subjective in the sense that there are no objectively correct answers to moral questions.

This analysis relates to AI because, as discussed below, it is not clear that AI is capable of attaining a subjective perspective. If the subjective perspective is required for moral judgment, then morality may be something AI cannot discern. And if legal or judicial practice at times requires moral judgment, then the role of AI in law may be circumscribed. All of that may be true—at least in theory—whether moral claims are objective or subjective (or something else).

The Scientific Perspective. To situate these points, it is important to clarify a few terms. The first is the scientific perspective, or what one might more technically call a naturalized, materialist reductionism. The notion is that all

---

67. For a similar discussion about the different potential meanings of subjectivity, see, for example, SEARLE, MIND: A BRIEF INTRODUCTION, *supra* note 4, at 94-95.

68. For arguments in favor of a non-naturalized moral objectivity, see, for example, RUSS SHAFER-LANDAU, MORAL REALISM: A DEFENCE (2003). I understand Thomas Nagel also to suggest a possibility along these lines. *See* NAGEL, *supra* note 10, at 97-126 (chapter 5, "Value").

knowledge can be reduced to what science can assess in the world. Defining the outer bounds of that science is controversial. It would presumably include at least physics, chemistry, and biology, which, as science advances, might collapse into one another. Other fields of study, such as psychology, might not be fully compatible with a naturalized, materialist reductionism. To the extent psychology uses empirical methods, it may well qualify as a true science. Over time, it may reduce first to biology, then to chemistry, and then to physics. But talk of the ego, the id, and the superego might disappear—or at least be mapped to concepts in the harder sciences—and so might the notion of the subconscious.[69] This Essay will treat the scientific perspective as equivalent to a naturalized, materialist reductionism.

Subjective Experience and Morality. To say that moral knowledge requires subjective experience can be understood as meaning that morality is not directly discernable from the scientific perspective.[70] The claim is that physics, chemistry, biology, and any other hard sciences cannot assess moral propositions. According to this view, there are no moral entities out in the physical world to be detected. Something about subjective experience—about a conscious mind and perhaps even a mind capable of self-consciousness, cognition, or both—is necessary to exercise moral judgment.

Subjectively Discerned Objective Morality. The above discussion could lead to doubt about the existence of objective moral truth. That could come from combining the scientific perspective with a belief that a subjective perspective is necessary for moral judgments. One might then believe that precisely because science cannot detect moral facts, those "facts" cannot be objectively true. Reasoning along these lines has led some philosophers to doubt moral objectivity. But there is another possibility. It may be that the scientific perspective is capable of leading to truth about the physical world, but that there are other realms of knowledge and that morality is one of them. The criteria for truth in morality may be different from the criteria for truth in science. Yet there still may be objective moral truths. Recent philosophers have pursued views along these lines.[71] They believe in non-naturalized objective moral objectivity. The arguments for and against that position are

69. *See infra* text accompanying notes 114-19 (discussing a modular understanding of consciousness).

70. Although science may not be able to discern morality directly, it can detect the expressed beliefs and conduct of moral agents. That may prove important in assessing the limits of AI.

71. *See, e.g.*, SHAFER-LANDAU, *supra* note 68; THOMAS NAGEL, THE LAST WORD 101-25 (1997) (chapter 6, "Ethics"); Ronald Dworkin, *Objectivity and Truth: You Better Believe It*, 25 PHIL. & PUB. AFF. 87 (1996).

extensive, intricate, and complex. This Essay will not attempt to engage them, much less resolve them.[72]

A brief summary of the above discussion may be helpful. The scientific perspective—a naturalized, materialist reductionism—holds that knowledge is available only through the hard sciences. One challenge to that view asserts that moral knowledge requires a subjective perspective and that, even though it does, moral claims may be objectively true (or false). If a challenge to the scientific perspective along these lines proves persuasive, then reasoning within the confines of science may not be capable of assessing the truth *vel non* of moral propositions. Morality may be objective in the sense that there are right and wrong answers to moral questions but subjective in the sense that knowledge about morality is accessible only from a subjective perspective, not from the scientific perspective.

This possibility sets the stage for investigating a possible limitation on the role AI can play in law (and other disciplines that may require moral judgments). AI appears to operate within the scientific realm. It may not be capable of attaining the subjective perspective and, if it cannot, it may not be able to make moral judgments (and perhaps other value judgments). That may set an outer boundary on the role computers can play in legal and judicial practice.

III. Reasons to Believe Computers Cannot Achieve Subjectivity

Whether computers are capable of having subjective experiences is no trivial issue. Again, this Essay does not seek to resolve it. But it does explore reasons to doubt that AI is capable of subjectivity. Its strategy is to note two related points: first, as far as we know, computers and all they do seem to be explicable in purely materialist terms, and, second, a materialist, scientific view has not yet been able to capture key aspects of subjectivity. To be sure, it may be that computers may someday achieve subjective experience. But, if so, we have no account at present of the conditions that would have to obtain for them to do so. It may also be true that someday we will develop a

72. At least one point does seem worth noting for the reader inclined to dismiss non-naturalized moral objectivity out of hand. An argument that the scientific perspective exhausts objective knowledge would not seem able to rely just on the scientific perspective without question-begging. Some larger perspective is necessary that can assess different potential forms of knowledge. That creates at least a challenge for the claim that the only truths worthy of the name are scientific truths. If that proposition itself is true, it would appear to be a non-scientific truth. Put differently, the philosophical argument that provides the foundation for the scientific view would not itself seem to be purely scientific.

materialist, reductionist account that fully captures subjective experience.[73] But we have not done so yet. Or so this Part argues. It explores the ways in which a scientific account of subjective experience is incomplete. Assuming that computers remain restricted to the realm of science, they therefore may not be able to achieve subjective experience. To the extent moral judgment requires subjective experience, computers—and AI—would then seem unable to make full moral judgments. Further, first-person decision-making in law seems to require moral judgments. So AI may not be able to displace entirely human beings in legal and judicial practice.

Part III explores the above line of reasoning by reviewing various controversies in the philosophy of mind. The scientific perspective has put great pressure in particular on three traditional notions: consciousness, free will, and the unified self. In each area, this Part argues, the scientific perspective has not yet fully captured the first-person perspective, at least not in ways that are capable of performing first-person decision-making.

A. Consciousness as an Illusion

> *Cogito, ergo sum.*[74]
> *Je pense donc je suis.*[75]
> *I think therefore I am.*

A particularly provocative claim about consciousness is that it is an illusion. Various theorists have made versions of this claim, perhaps most notably Daniel Dennett.[76] There is controversy about what the claim entails—and about whether its proponents are consistent in characterizing the proposition. For present purposes, regardless of the positions theorists in fact

---

73. John Searle takes a particularly interesting position on these issues. He characterizes his position as naturalist but not as eliminative reductionist. In other words, he believes it will ultimately prove possible to explain first-person experiences as part of the natural world and that it will not prove necessary to deny the existence of first-person experiences to maintain our commitment to science. *See, e.g.*, SEARLE, MIND: A BRIEF INTRODUCTION, *supra* note 4, at 79-80. He labels his position "biological naturalism." *Id*. at 79. The notion is that consciousness is a "higher-level" feature—or an "emergent quality"—of the biological brain. *Id.* at 79-80; SEARLE, MIND, LANGUAGE, AND SOCIETY, *supra* note 57, at 52-54; JOHN SEARLE, THE MYSTERY OF CONSCIOUSNESS 8, 13, 161, 210-14 (1997) [hereinafter SEARLE, MYSTERY OF CONSCIOUSNESS]. A full discussion of Searle's position and its implications is beyond the scope of this Essay.

74. RENATI DES-CARTES [RENE DESCARTES], PRINCIPIA PHILOSOPHIÆ pt. I, § 7, at 2 (1644), https://www.wdl.org/en/item/3157/view/1/29/.

75. RENE DESCARTES, DISCOURS DE LA METHODE 32 (Paris, Leopold Cerf, 1902) (1637), https://zulu-ebooks.com/send/3-fachbuecher/859-discours-de-la-methode.

76. *E.g.*, DANIEL C. DENNETT, CONSCIOUSNESS EXPLAINED 309-14 (1991).

adopt, it is perhaps most useful to start with the claim in its most straightforward form: we do not really have consciousness; we just think we do.

If consciousness does not really occur—if we only think it does, whatever that means—then the first-person perspective too would seem to be an illusion. Consciousness is constitutive of a person's perspective. Without consciousness, there is no first-person perspective and hence no need for science to capture it.

But the claim that consciousness is an illusion seems self-defeating. An illusion—of all things—is a creature of perspective. It is an instance in which subjective experience conflicts with reality. Without subjective experience, no such conflict is possible. Without conscious beings, no one is capable of experiencing an illusion.

Moreover, consciousness is the one aspect of the world that we experience directly. Careful analysis of anything else we think we know about reality shows how tenuous our knowledge is.[77] This is not the place to explore longstanding debates about the extent to which we impose frameworks of knowledge on the world as opposed to derive those frameworks from the world. But those debates are vexing enough for many to find wisdom in Descartes' famous pronouncement, "I think therefore I am." Our conscious experiences provide us a thin reed of knowledge in the face of radical doubt. Our ability to doubt arguably provides us the most reliable evidence we have about the world—that, if nothing else, there is a self capable of doubting.[78]

Why then deny consciousness exists? One answer is that it provides a way to sidestep a profound problem that has yet to be solved: the relationship between the physical world and conscious existence. The difficulty in this regard is apparent in the debate over a famous thought experiment by the philosopher John Searle: the Chinese Room. Searle asks you to imagine you are in a locked room. You understand only English. You receive various written materials in Chinese. You have various sets of instructions in English for correlating the batches of Chinese texts, relying only on their shapes. By following the instructions, you are able to conduct a "conversation" in Chinese about stories written in Chinese, responding to questions in a way that would fool a native speaker. Searle contends that you still do not understand Chinese merely because you are able—through the instructions—to simulate an understanding of Chinese. You are just "manipulating

77. For a classic argument, see GEORGE BERKELEY, TREATISE CONCERNING THE PRINCIPLES OF HUMAN KNOWLEDGE (1710).

78. *See* NAGEL, *supra* note 10, at 82.

uninterpreted formal symbols."[79] He reasons that the same is true, in effect, for computers. They can be built to appear to understand language but that does not necessarily mean that they really understand, not the way human beings do. This thought experiment has created quite a stir. It provoked various responses. Searle has replied to many of them.[80] Neither side seems persuaded by the other.

At the core of this persistent disagreement over Searle's thought experiment arguably lies our ignorance about how biology gives rise to a conscious mind. As a result, one might say that how persuasive the thought experiment is depends on where one begins one's analysis. On one hand, it does seem obvious—as Searle argues—that a more complicated version of his thought experiment would not necessarily change the result. Mere manipulation of symbols is not enough for understanding—or relatedly, as Searle puts it, syntax does not constitute or suffice for semantics.[81]

On the other hand, it is not clear what physical states give rise to a conscious mind—or what needs to be added to syntax to constitute and suffice for semantics. How does human physiology—including the brain and other parts of our physical being—differ from complicated computers? One might think of human beings and computers as consisting of two parts. Call them hardware and software or substrate and pattern.[82] What, if anything, is so special about our biological form that can sustain a mind in a way that a computer cannot? Many modern thinkers believe ultimately in a materialist naturalism that leads to reductionism. For them, it would seem, the answer is nothing—nothing is so special about our biology that precludes a computer

79. John R. Searle, *Minds, Brains, and Programs*, 3 BEHAV. & BRAIN SCI. 417, 418 (1980).

80. For a useful summary of the literature and arguments, see Josef Moural, *The Chinese Room Argument in* CONTEMPORARY PHILOSOPHY IN FOCUS: JOHN SEARLE 214-60 (Barry Smith ed., 2003).

81. It is important to be clear about Searle's argument. He does not deny that computers may at some point become conscious. His position instead is that just because a computer can simulate consciousness does not necessarily mean it is conscious. SEARLE, MYSTERY OF CONSCIOUSNESS, *supra* note 73, at 13-15, 110. He rejects in particular the notion that the mind is merely a computer, and the apparent corollary that any computer that can simulate consciousness has achieved consciousness. *See* SEARLE, MIND: A BRIEF INTRODUCTION, *supra* note 4, at 46-52 (explaining the computer theory of the mind); *id.* at 58-73 (presenting arguments against the computer theory of the mind and other forms of materialism and responding to potential counterarguments).

82. *See* TEGMARK, *supra* note 1, at 24-30 (discussing hardware and software); *id.* at 65-67 (discussing substrate-independent patterns).

from supporting a conscious mind.[83] From that perspective, the Chinese Room experiment is just a clever way to require us to identify when the ghost enters the machine, even though we have no satisfactory account of what the ghost is or how it can exist at all.

Calling consciousness an illusion may just be a way to sidestep this difficulty. If we cannot reconcile our scientific account of the physical world and conscious experience—if we cannot build an equation that includes both—one solution would be to deny that we need such an equation. That view is manifest in Searle's claim that denying consciousness is not a solution to the hard problem but a way of avoiding it. Such an approach "changes the subject. It is not about consciousness, but rather a third-person account of external behavior."[84]

For present purposes we need not resolve this debate (thank goodness!). Two more modest points are relevant: first, whatever the merits of denying consciousness, doing so is unlikely to provide guidance for the first-person perspective; and, second, even if denying consciousness entirely does not advance first-person thinking or decision-making, some insights from science about how the mind works may do so. Each of these points warrants a bit of elaboration.

Let us begin with the first point: whatever the best third-person account of consciousness is—that it is an illusion or something else—the claim that consciousness is an illusion is unhelpful from a first-person perspective. Accepting that consciousness is not real does not provide guidance for our (apparent) conscious perspective. When someone sits down to mull a decision—whether it is about what to do on a rainy Saturday or what the best interpretation of a legal precedent is—the notion that the person is not really conscious does not seem capable of advancing the effort at all. Someone may think she is thinking and yet be mistaken, but if she is mistaken, how can she proceed differently? What alternative does she have to engaging in what feels like a conscious process for arriving at the best decision she can make?

That first point may lead us to conclude that scientific insights cannot inform first-person thinking at all. But doing so could be imprudent. While our first-person perspective likely cannot abandon the notion of consciousness—our first-person perspective would not benefit from denying its own existence—more modest scientific claims may have practical value for the first-person perspective. Put differently, even if consciousness is not

83. For a fascinating discussion of these issues, with an iconoclastic skepticism about materialistic naturalism and its associated reductionism, see NAGEL, *supra* note 10.

84. SEARLE, MYSTERY OF CONSCIOUSNESS, *supra* note 73, at 123 (replying to Daniel C. Dennett).

just an illusion, some of our conscious experience may consist of illusions. Recognizing that is so may lead to insights that can improve our first-person experiences and decision-making, even if they do not cause us to abandon the notion of consciousness entirely. This last point finds useful application to a related proposition: that free will is an illusion.

B. Free Will as an Illusion

> *We must believe in free will. We have no choice.*[85]
> *Don't believe everything you think.*[86]

A similar analysis applies to the related—and fraught—concept of free will. Philosophers have developed fascinating and complex ways of understanding the nature of free will, determinism, and moral responsibility, and the relationship between them. These are worth canvassing, at least briefly. But what matters most for present purposes is that the philosophical debates have focused on the third-person perspective—about the best objective understanding of free will—and not on the first-person perspective—what is useful for first-person decision-making. As a result, arguably much of the debate cannot helpfully inform first-person decision-making—much, but not all. There may be some knowledge and insights from the literature on free will than can help people—including lawyers and judges engaging in legal and judicial practice—in making decisions.

To frame this discussion, we should define some key terms. Unfortunately, there is no uncontroversial definition of free will. Nor is there an uncontroversial way to characterize the relationship between free will, determinism, and moral responsibility. That said, it is useful to have at least some working notions in mind. One way to define free will is as requiring "alternative possibilities or the power to do otherwise."[87] In other words, according to this definition, a person has free will if she could have acted differently than she did—she had the power to choose an alternative course of action. A second way to define free will is that the person is the "ultimate source" of her free actions—or she is the ultimate source of her will to perform free actions.[88] Note that these two definitions are not the same. A

85. Stefan Kanfer, *Isaac Singer's Promised City*, CITY J., Summer 1997, https://www.city-journal.org/html/isaac-singer%E2%80%99s-promised-city-11935.html.

86. CafePress Bumper Sticker, https://www.amazon.com (search query: "CafePress Don't Believe Everything You Think").

87. *E.g.*, JOHN MARTIN FISCHER ET AL., FOUR VIEWS ON FREE WILL 1 (2007) [hereinafter FOUR VIEWS ON FREE WILL].

88. *Id.*

person, for example, might have been able to take only a single relevant action, given who she is and her environment, but she might nonetheless be understood as having acted on her own free will if she engaged in the appropriate kind of decision-making, unconstrained, for example, by certain kinds of internal pressures (*e.g.*, a brain tumor) and external pressures (*e.g.*, a gun to the head).

In categorizing the different philosophical schools of thought on free will, it is useful to ask of each whether its adherents believe that free will is compatible with determinism. Determinism can be defined as requiring "that at any time the universe has [only] one physically possible future."[89] So-called "incompatibilists" believe that free will and determinism are not compatible, and "compatibilists" believe that they are. One major view within incompatibilism is held by libertarians, who believe we do have free will and reject determinism. Another such major view is held by hard incompatibilists, who believe that we lack free will (and may or may not believe the world is deterministic). Incompatibilists generally agree that determinism is incompatible with moral responsibility, so that libertarians believe that we do have moral responsibility for our actions and hard incompatibilists believe that we do not. Compatibilism takes many forms, and many compatibilists believe that moral responsibility is consistent with determinism.[90]

The arguments made by philosophers in each of these schools are subtle and complex, as are the variations on the proposed solutions to the problem of free will. Most of them have no obvious bearing on the issues before us. In particular, if we reflect on our own experiences as first-person decision-makers, we are likely to offer an account of free will that is rarely held among philosophers: a commonsense libertarianism. We experience ourselves as having a range of options available to us, as being free to select among them, and as making a choice through an exercise in agency that is affected but not determined by our environment or our physical constitution (or by any chance occurrences that may happen in one or the other). Further, we experience our process of deliberation as determining the choices that we make. If we were to reason differently—or to abandon reason in favor of our intuitions or "gut instincts" or vice versa—we might act in some other manner than we do.

89. *Id.* at 2.
90. *Id.* at 4.

Sam Harris—a hard incompatibilist—captures these experiences nicely. He writes:

> Our moral intuitions and sense of personal agency are anchored to a felt sense that we are the *conscious source* of our thoughts and actions. When deciding whom to marry or which book to read, we do not feel compelled by prior events over which we have no control. The freedom that we presume for ourselves and readily attribute to others is felt to slip the influence of impersonal background causes.[91]

In other words, we experience ourselves as having a particularly free version of free will. To paraphrase Aristotle, we believe ourselves to be unmoved movers.[92]

Contrast this account with what one might think of as a sophisticated libertarianism. Robert Kane offers a thoughtful philosophical argument for libertarianism. But to make it robust—to make it stand up to sophisticated critics—his account strays far from the actual experience of decision-making that we likely find familiar. He is skeptical, for example, of the claim that there is a source of action within us that cannot be explained in traditional scientific—and empirical—terms.[93] To make room for libertarianism, then, he concedes that most decision-making may not be consistent with free will. And he attempts to provide a materialist account of free will, one based in brain science.

Toward that end, he identifies conduct at key inflection points—which he calls "undetermined self-forming actions"—that can change the course of our lives.[94] He offers as an example a businesswoman who is on her way to an important meeting when she witnesses an assault in an alley, forcing her to choose between her conscience—her desire to help the victim—and her

---

91. SAM HARRIS, FREE WILL 16-17 (2012).

92. For arguments that we cannot help but experience free will in this way see, for example, DWORKIN, JUSTICE FOR HEDGEHOGS, *supra* note 16, at 224 ("In the first person, deciding includes assuming judgmental responsibility; the connection is internal and independent of any premise about the causes of decision. Pessimistic noncompatabilism is not an intellectually stable position. It asks us to believe what we cannot believe."); SEARLE, MIND: A BRIEF INTRODUCTION, *supra* note 4, at 153-54 ("Whenever we decide or act voluntarily, which we do throughout the day, we have to decide or act on the presupposition of our own freedom. Our deciding and acting are unintelligible to us otherwise. We cannot think away our own free will.").

93. Robert Kane, *Libertarianism in* FOUR VIEWS ON FREE WILL, *supra* note 81, at 5, 24-25.

94. *Id.* at 26.

ambition—her commitment to attend the meeting.[95] He then combines parallel processing in the brain with chaos theory and quantum mechanics to argue that on these unusual—but important, indeed formative—occasions, it is possible that people "*make* one set of competing reasons or motives prevail over the others then and there *by deciding*."[96] Whatever the merits of this approach as a third-person account of free will,[97] it seems to cede too much to non-libertarianism to capture the first-person experiences of most people as they make many routine conscious decisions. Harris's commonsense description of libertarianism provides a more familiar description of our first-person experience. It reflects our decision-making process as we ordinarily understand it from the inside.

Harris also makes another important observation: that the deliberative process matters (although he believes it is the product of other forces and chance):

> And the fact that our choices depend on prior causes does not mean that they don't matter. If I had not decided to write this book, it wouldn't have written itself. My choice to write it was unquestionably the primary cause of its coming into being. Decisions, intentions, efforts, goals, willpower, etc., are causal states of the brain, leading to specific behaviors, and behaviors lead to outcomes in the world. Human choice, therefore, is as important as fanciers of free will believe. But the next choice you make will come out of the darkness of prior causes that you, the conscious witness of your experience, did not bring into being.[98]

The last sentence gives the reader a fuller and fairer sense of Harris's view, but it is the rest of the quotation that matters for the present discussion. Harris acknowledges the importance of decisions, intentions, efforts, goals, and even will power. They are, as he says, causal states of the brain. And presumably we would not necessarily take the same actions we do if we did

95. *Id.*

96. *Id.* at 26-27.

97. It is a fair question whether Kane has provided an adequate alternative to the notion of "agent-causation," which he rejects, *id.* at 25, or whether instead he has relegated it to a small but crucial role. In other words, one might doubt that he has provided an adequate account of how the physical world can involve actions other than those that are predetermined or produced by chance events (quantum mechanics) without appealing to the notion of agent-causation. But addressing that issue is beyond the scope of this Essay.

98. HARRIS, *supra* note 91, at 34.

not engage in our ordinary deliberative process, the one that feels like it is driven by our uncaused causal agency (which Harris believes is illusory).[99]

Harris's view is that we would do better to accept determinism and abandon notions of free will and moral responsibility. We should not blame people for the actions they take and the harms they cause, any more than we would blame a dog for attacking someone or, if you prefer, a bolt of lightning for striking someone. We should instead ask a prospective and practical question: what steps should we take to prevent undesirable future harms? There is much to say in favor of Harris's view and much to say against it. But the point here is not to assess whether Harris's philosophical argument for determinism is persuasive, and so there is no need to explore it further. His focus is on a third-person perspective—how we should evaluate human behavior from the outside—and the inquiry before us is about the first-person perspective.

Harris has some useful points to make about the first-person perspective as well. As noted above, he recognizes how we ordinarily experience first-person decision-making. He also suggests a deterministic first-person alternative. He touches on it only briefly, as his focus is on third-person truth. In particular, Harris ends his book on free will by attempting to offer a first-person account of determinism. It is hard to describe it other than as striking a false note. Consider how he describes his decision to conclude his book, *Free Will*:

> What brings my deliberations on these matters to a close? This book must end sometime—and now I want to get something to eat. Am I free to resist the feeling? Well, yes, in the sense that no one is going to force me at gunpoint to eat—but I am hungry. Can I resist this feeling a moment longer? Yes, of course—and for an indeterminate number of moments thereafter. But I don't know why I make the effort in this instance and not in others. And why do my efforts cease precisely when they do? Now I feel that it really is time for me to leave. I'm hungry, yes, but it also seems that I've made my point. In fact, I can't think of anything else to say on the subject. And where is the freedom in *that*?[100]

In an otherwise insightful analysis, what stands out as particularly implausible is Harris's description of why he stopped writing his book when

99. *Id.* at 5 ("Free will *is* an illusion. Our wills are simply not of our own making. Thoughts and intentions emerge from background causes of which we are unaware and over which we exert no conscious control. We do not have the freedom we think we have.").

100. *Id.* at 66.

he did. It does not seem credible that he ended the book because he was hungry. Did he write the entire tome one day between breakfast and lunch? Writing—and rewriting—is a highly deliberative process. It involves careful thought, revision, reflection, and methodical action. Why would hunger pangs cause him to alter his plans?

Nor is it plausible that, spur of the moment, he could not think of anything else to say. His short book reveals an efficiency and restraint inconsistent with an extemporaneous decision to stop writing and grab a bite to eat. Presumably, he organized his thoughts in an outline—formal or informal, in his head, on a screen, or on a notepad—and worked through it over many days, if not weeks or months, reviewing, revising, and mulling as he went along. Indeed, as discussed above, he acknowledges that deliberation plays a causal role in decision-making.[101] So his implication that writing a book is a sort of obviously mysterious process—with conscious decision-making playing little to no role—does not ring true.[102]

That said, Harris may be right that mystery shrouds the ultimate sources of preferences—why we like chocolate ice cream more than vanilla ice cream or ice skating more than basketball (if we prefer skating because we think artistry for its own sake is more pleasing than artistry in the service of a concrete goal, then why do we hold that view?).[103] From a first-person perspective, however, we have significant control over how much of our decision-making process is based on something that approximates an

---

101. John Searle takes a similar position on the causal role of consciousness in human action, one for which he offers a particularly intriguing philosophical argument. *See, e.g.*, SEARLE, MIND, LANGUAGE, AND SOCIETY, *supra* note 57, at 58-62, 104-07; SEARLE, MIND: A BRIEF INTRODUCTION, *supra* note 4, at 91, 136-50.

102. Harris elsewhere recognizes this distinction. *See, e.g.*, HARRIS, *supra* note 91, at 32-33 ("This is not to say that conscious awareness and deliberative thinking serve no purpose. Indeed, much of our behavior depends on them. I might unconsciously shift in my seat, but I cannot unconsciously decide that the pain in my back warrants a trip to a physical therapist. To do the latter, I must become aware of the pain and be consciously motivated to do something about it. Perhaps it would be possible to build an insentient robot capable of these states—but in our case, certain behavior seems to require the presence of conscious thoughts. And we know that the brain systems that allow us to reflect upon our experience are different from those involved when we automatically react to stimuli. So consciousness in this sense, is not inconsequential.").

103. *Id.* at 39 ("Choices, efforts, intentions, and reasoning influence our behavior—but they are themselves part of a chain of causes that precede conscious awareness and over which we exert no ultimate control. My choices matter—and there are paths toward making wiser ones—but I cannot choose what I choose. And if it ever appears that I do—for instance, after going back and forth between two options—I do not *choose* to choose what I choose. There is a regress here that always ends in darkness.").

automatic reaction and how much we subject to a more careful, deliberative process, even if the ultimate source of our beliefs, feelings, intuitions, and the like may remain obscure to us. From a first-person perspective, we experience ourselves as having a choice about how careful we are in making the decisions that we make.[104] In that sense, whatever the merits of determinism[105] as a third-person account of human decision-making, it fails to capture our first-person experience.

Here science may inform our understanding of our options. The first-person perspective can accommodate—and perhaps benefit from—distinguishing between the mindlessly automatic and the mindfully reflective. Consider Daniel Kahneman's famous distinction between fast and slow thinking.[106] Fast thinking is quick, automatic, easy, based on heuristics, and prone to various cognitive biases. Slow thinking, in contrast, is sluggish, deliberate, and difficult, and can be logical and can overcome cognitive biases.

Consider also a fascinating set of experiments by Benjamin Libet that bear on free will. Subjects were asked to choose when to flex their wrists.[107] They experienced making a conscious choice to do so. But monitoring of their brains suggested that the initial brain activity associated with a physical movement occurred well before the subjects experienced a conscious decision to act.[108] The conscious "decision," then, may have been a *post hoc*—and non-causal—step in a process that was in fact initiated in a non-conscious manner. Later experiments—by Libet and others—appear to confirm this timing.[109] But Libet did additional experiments suggesting that while certain—apparently unconscious—brain activity was a necessary predicate for movement, people were able to make a conscious decision later in the process *not* to move. This has been cleverly put as showing that while we may not have free will, we have "free won't."[110] The parallel to Kahneman's model of fast and slow thinking is intriguing. Perhaps people,

104. I am not claiming that Harris denies this; to the contrary, I understand him to acknowledge it.

105. By determinism here, I do not mean to exclude a possible role for chance, such as is implied by quantum mechanics. *Id.* at 27-30.

106. DANIEL KAHNEMAN, THINKING, FAST AND SLOW (2013).

107. Benjamin Libet, *Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action*, 8 BEHAV. & BRAIN SCI. 529, 529-30 (1985).

108. *Id.* at 536.

109. *See, e.g.*, SUSAN BLACKMORE & EMILY T. TROSCIANKO, CONSCIOUSNESS: AN INTRODUCTION 231 (3d ed. 2018) (citing M. Shultze-Kraft et al., *The Point of No Return in Vetoing Self-Initiated Movements*, 113 PROC. NAT'L ACAD. SCI. 1080 (2016)).

110. *Id.* (quoting a personal communication from Richard Gregory).

from the first-person perspective, can impose conscious decisions to act—or not to act—on their non-conscious reactions, leading to more intentional and hopefully wiser behavior.

None of this necessarily falsifies the claim—by Harris and others—that free will understood in the commonsense libertarian way does not exist. But it does seem to create space from within the first-person perspective for distinguishing decisions made in an automatic and thoughtless manner from those that are deliberate and conscious. And that distinction may preserve a realm for what human beings are uniquely capable of doing—and what computers may not be capable of doing.

This discussion of free will, then, tends to support a few relevant points. First, the third-person, scientific perspective fails to capture in important ways the first-person experience of free will. Second, and related, the scientific, reductionist perspective has limited utility for the first-person perspective, particularly when it comes to decision-making. Third, some insights from science may usefully inform first-person decision-making, but only when they are relatively modest and do not purport to displace free will entirely. Similar points apply to the notion of a unified self.

C. The Unified Self as an Illusion

> *There was a faith healer from Deal*
> *Who said 'though I know pain isn't real.*
> *If I sit on a pin,*
> *And it punctures my skin,*
> *I dislike what I fancy I feel.*

1. The Singular Self at a Given Time

Another way in which a scientific, third-person account of consciousness appears to deviate from our first-person experience involves the unified self. We generally experience ourselves as having a singular conscious existence. I may have mixed motives or ambivalence about some matters, but I am only one person, not two or more people occupying a single body and competing with one another, negotiating internal compromises, and the like.

There may be exceptions. We are fascinated, for example, by people who suffer from what is commonly called multiple personality disorder, and more formally known as dissociative identity disorder or "DID."[111] People with multiple personalities do not seem to have a single identity but rather multiple

111. AM. PSYCHIATRIC ASS'N, DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS code 300.14 (5th ed. 2013) (under "Dissociative Disorders").

identities vying for control and surfacing at different times. At times DID has captured the popular imagination, including in shows such as *The United States of Tara*. But we think of DID as an aberration and foreign to our ordinary experience. Indeed, that is in part why we are fascinated by it. It seems unreal, like magic or superpowers. Unsurprisingly, DID is a source of controversy within the psychiatric profession. Ordinary people may find it so hard to relate to people with multiple personalities that a natural response is to believe they are pretending.

It is therefore all the more surprising—and fascinating—that science has suggested a model for consciousness that in some ways is similar to DID.[112] One way to get at this point is to consider what Daniel Dennett has usefully called the "Cartesian Theatre."[113] Many of us have a sense that there is a unified self that has various experiences. It takes in information from our five senses and in response has various feelings and thoughts and makes various decisions. But scientists suggest that may not be how our brains work.

The difficulty lies in part in that there is no identifiable part of the brain that serves as a kind of control center or headquarters. Instead, there appear to be different portions of the brain that engage in parallel processing and that may have a larger or smaller role in causing particular behaviors at a given time. One useful way to think of the brain is as containing different "modules."[114] Robert Wright provides a particularly accessible explanation of how to understand this idea. The notion is that different parts of the brain have evolved through natural selection—and chance—to perform different functions, and depending on the circumstances different modules may be in ascendancy at particular times.[115] Wright usefully cautions us not to think of modules as discrete.[116] They are not separate units, sealed off from one another. With that caution in mind, the analogy is helpful. In a sense, it suggests that we are all a bit like people who suffer from DID. Different parts of our brain compete for control, and which one prevails at a given moment may dictate what we think, believe, and do. And there is no dominant, overarching self that selects among the competing modules.

---

112. I do not mean to imply any direct correlation between the brain science I discuss in this Section and DID. There may or may not be one, but I mean to take no position on that issue.

113. DENNETT, *supra* note 76, at 107.

114. ROBERT WRIGHT, WHY BUDDHISM IS TRUE: THE SCIENCE AND PHILOSOPHY OF ENLIGHTENMENT 94 (2017).

115. *Id.*

116. *Id.*

Whatever the merits of a third-person, "modular" view of brain functioning—it is unsurprisingly controversial[117]—it seems to have limited utility from the first-person perspective, at least in a strong form. If there is no unified self—no Cartesian Theatre—from which we can view the different modules, but rather only a series of different times at which particular modules dominate, what can we do with that information? It seems to hold (potentially) great descriptive and predictive value, but does not offer any obvious route for improving our experience or our decision-making. And it is so foreign from our first-person perspective that there is no clear way to synthesize the two.[118] The modular theory denies the existence of a unified self capable of undergoing experiences, contemplating options, and initiating actions. So what self could act on the knowledge of modularity?[119]

Wright, in his thoughtful and provocative book on cognitive science and Buddhism, acknowledges this issue. Some traditional understandings of Buddhist doctrine, much like some modern scientific understandings, deny the existence of any self at all. Yet Wright is committed to incorporating insights from both Buddhism and science in a practical prescription for how to improve how we think and live. His proposal is that we use meditation to obtain a certain level of remove from our own senses, emotions, thoughts, and beliefs. From this perspective, we may be able, as it were, to watch our modules in action—and to view what they cause us to experience, feel, think, and believe from a critical distance. That separation, according to Wright, may allow us to make calm, informed, and deliberate judgments about who we want to be and what we want to do. We gain some control over ourselves.

Note that Wright's description of how the mind works and his prescription for how we might respond to that information do hold great potential value from the first-person perspective. He suggests a way for us to gain greater mastery over ourselves and to exercise our wills in ways that may not come naturally to us. But, as he recognizes, there is a difficulty in reconciling his

---

117. *Compare* JEROME H. BARKOW ET AL., THE ADAPTED MIND: EVOLUTIONARY PSYCHOLOGY AND THE GENERATION OF CULTURE (1992) *and* JEFF CLUNE ET AL., THE EVOLUTIONARY ORIGINS OF MODULARITY, PROCEEDINGS OF THE ROYAL SOCIETY, Mar. 2013 (arguing for a modular theory), *with* Jaak Panksepp & Jules B. Panksepp, *The Seven Sins of Evolutionary Psychology*, 6 EVOLUTION AND COGNITION 108, 108-09 (2000) (questioning the modular theory).

118. *See* SEARLE, MIND: A BRIEF INTRODUCTION, *supra* note 4, at 200-06.

119. It is interesting to consider how the modular theory can account for knowledge about the modular theory. How can a person know his or her own brain is modular? Presumably one or more modules contain this knowledge and use it as part of the ongoing struggle among the modules for supremacy.

argument with the notion that there is no self.[120] To act as Wright suggests, we seem to have to accept the idea of a Cartesian Theatre or something similar to it. Otherwise, who is it that, after the benefits of meditation, is able to make better choices and to exercise more control?

Again, two propositions seem to be true: the third-person, scientific perspective in its pure form seems incompatible with—and unhelpful for—first-person decision-making; and yet some of the insights we gain from science may be helpful to that decision-making, if appropriately adapted.

2. The Singular Self over Time

In addition to casting doubt on the existence of a single self at a given time, the scientific perspective casts doubt on the persistence of the self over time. In other words, the person I am right now and the person I am in a minute (or a moment) from now may be in a profound sense disconnected, despite our intuitive sense of our own continuing existence. From the first-person perspective, this notion seems preposterous. From the third-person perspective, it is surprisingly hard to resist. But even if true, the surprising—even startling—idea that we in effect die in every moment seems difficult to integrate in a useful way into the first-person perspective.

Consider the concept of whole brain emulation (WBE), also known as mind uploading or brain uploading. A startup already exists, Nectome, that

120. WRIGHT, *supra* note 114, at 64. Wright recognizes this paradox, although he does not necessarily take the position, suggested above, that a self must exist to render his prescription viable. A similar difficulty—a potential internal inconsistency—besets Susan Blackstone's attempt to reconcile her view that there is no self in the ordinary sense with how human beings see the world. She concludes her fascinating short book summarizing issues that arise about consciousness as follows:

> My hope is that one day our scientific understanding of consciousness will come together with personal insight. There are already some scientists with deep personal practice, and practitioners who study the science, holding out the hope that first- and third-person perspectives will eventually come together and let us see clearly. Both intellectually and in our own experience we should be able to stop being deluded and see through all those illusions of self, free will, and consciousness.

SUSAN BLACKMORE, CONSCIOUSNESS: A VERY SHORT INTRODUCTION 133-34 (2d ed. 2017). What Blackmore does not explain—if self, free will, and consciousness are illusions—is who the "we" is that will gain this clarity, how we can choose to pursue it, and how we can stop being deluded—or how we can ever be deluded in the first place—if consciousness is an illusion.

plans to offer this service to terminal patients.[121] It offers people the prospect of achieving immortality by uploading their minds to a computer. Assume for the moment this is possible. Assume that technology advances such that a computer can replicate biological existence. People may be able to see, smell, feel (in the physical and emotional senses of that word), think, believe, etc., much as they did when they had an organic form, or at least close enough that their existence in computer form could feel, from their new perspective, continuous with their past existence. Not just their capacity to process current experiences, but also their memories, are uploaded and stored, so that the computer versions of people may feel that they are the very same people living a continuous life with their former carbon-based selves. Assume that the computer versions can last forever through replacement of hardware over time. Will they have achieved immortality?

You might think they will not have. When their human form dies, *they* may have died, and they may have merely created immortal replicas of themselves. That seems most consistent with our first-person perspective. Assume, for example, that someone's human self survives the uploading intact. The organic version and the computer version co-exist for some period. The two may even interact. In our ordinary understanding of these matters, we would say that the organic form remains who a person is and that the computer form is someone else (assuming it is a person at all). The actual organic person would likely perceive the computer facsimile as a separate entity, like a twin, only far closer to identical in certain ways than even an identical sibling.

This issue is not new to philosophy, although the particular application—uploading a mind to a computer—is relatively novel as a feasible possibility, if it is one. Philosophers have long asked what, if anything, about the self endures over time. Consider a hypothetical famously discussed by Derek Parfit, among others. Parfit asked us to imagine that it becomes possible to transport someone to a distant location through a scanning process, perfectly reconstructing the person on a lunar colony and then destroying the original version.[122] Assuming that the memories of the original version of the person are included in the new version—presumably memories exist in some physical form within us—the reconstructed person experiences no discontinuity, other than a sudden change of location. Would you step into the scanner? If you did, would you survive the trip?

---

121. Antonio Regalado, *A Startup Is Pitching a Mind-Uploading Service That Is "100 Percent Fatal*," MIT TECH. REV. (Mar. 13, 2018), https://www.technologyreview.com/s/610456/a-startup-is-pitching-a-mind-uploading-service-that-is-100-percent-fatal/.

122. DEREK PARFIT, REASONS AND PERSONS 199 (1987).

Parfit puts a finer point on the problem by proposing a variation: the original version of you is not destroyed at the time of the scanning. You survive in your original form. So two versions of you exist, one far away (this time on Mars) and one on Earth. But the scanning has caused a problem for the Earth version of you. It will die within a few days.[123] Do you, as the Earth version, take any comfort from the continued existence of a version of you on Mars? More precisely, when your Earth version dies will *you* continue to exist? You may not. You may well feel as much panic about dying imminently as you would if you had not been scanned. It is well and good for the replica of you that it will continue to exist, you might think, but that does not do the *real* you any good at all.

What is intriguing—and potentially upsetting—is how hard it is to resist the view from a scientific, materialist perspective that our ordinary lives are in key respects just like the person who is uploaded to a computer and then dies or the person who is scanned and destroyed. The reason is that from the scientific view it is hard to account for any persistent self at all. To be sure, science can explain our *sense* of continued existence. The human mind is configured in such a way that it believes in and highly values its persistent existence. That is highly adaptive from the perspective of evolution. Creatures who understand themselves as persisting as a self over time and who prize their continued existence—likely above almost all else—are likely to survive long enough to procreate. But from a scientific perspective, the most plausible account of our sense of an enduring self is just a delusion. We have only present experiences—pains, pleasures, etc.—combined with memories of past experiences—also presently experienced—and a psychology built to think of ourselves as enduring over time. But two entities that have the same—or equivalent—present experiences and memories and psychologies would from a purely scientific, third-person perspective seem to have equal claim to be the same self.[124]

To get at least some sense of the strength of this position from the scientific, third-person perspective, imagine that the scanning works as follows. You enter a dark box and are perfectly duplicated. Two organic versions of you emerge. No one knows—not even you or the duplicate of you (if the other version is the duplicate)—which one of you is the original. You both have the very same memories, including the memory of entering the box and exiting it. You both have the very same psychologies. You have equivalent—if ever so slightly different—experiences of leaving the box.

123. *Id.* at 199-200.

124. *See* SIMON BLACKBURN, THINK 120-48 (1999).

Assume even the age of the physical materials that constitute your forms are indistinguishable, such that science cannot tell which one of you was scanned and which one created from the scanning process. The new version of you would feel the very same sense of the persistent existence of the self as the original version.

This scientific view of how the self endures over time—or does not endure over time—is known as the bundle theory. The label dates back to Hume, who wrote that he could identify no persistent self over time, just a bundle of experiences.[125] The most powerful counterarguments to the bundle theory tend to rely on concepts that are generally considered foreign to the scientific account, such as the notion of a spirit or a soul. If a spirit or soul does exist, then presumably it remains in the original version of you and not the replica produced through the scanning process. But if there is no spirit or soul or any equivalent, it is hard to say why the original you has any stronger claim to be you than the copy does.

Incorporating the bundle theory into the first-person perspective could lead to some profound shifts in how we see the world and behave. The consequences would seem to be great if we were to treat our future selves in significant ways as not really the same as our current selves—as people might think of a scanned or computer replica of themselves out there in the world. It would seem to imply, for example, that the future pains or pleasures that I will experience are not really going to be experienced by me any more than the pains or pleasures of a replica or, for that matter, someone completely unrelated to me. That might incline us toward the utilitarian view—which many people find counterintuitive—that we should treat everyone's pain and pleasure equally and not privilege our own experiences over those of others. Perhaps the utilitarians are right because we are not the same person from moment to moment. Perhaps teenagers are too when they act recklessly—not because they are immortal, as they seem to think, but because we are all perfectly mortal, dying in each instant with only in effect an unconnected replica existing in the next one. This notion brings new salience to the myth of the phoenix rising from its ashes. And perhaps Robert Wright should have included in his discussion of the connections between Buddhism and modern science an exploration of the idea that all that exists is the present.

Unlike the idea of psychological modules, however, most of the implications of the bundle theory for decision-making are not obvious. On

125. See BLACKMORE & TROSCIANKO, *supra* note 109, at 437-38 (citing PARFIT, *supra* note 122). Some strands of Buddhism have long held that there is no persistent self over time. *Id.* at 436. One can understand this doctrine as having important similarities to Hume's view. *Id.* at 436-38.

one hand, the notion that we are perpetually and persistently dying would seem to be profoundly disruptive for our first-person perspective. If we were to accept this proposition, it would seem we move through the world in a manner very different than the way we currently do. On the other hand, it is not clear how it would change the way we see the world and make decisions.

D. Summary: Science and the First-Person Perspective

The above analysis is meant to be suggestive. It explores reasons to believe that science is not currently capable—and may not ever be capable—of providing a complete account of subjective experience. To the extent computers operate in the realm of science—to the extent that they are able to process information only in ways that science can explain—they may not be able to attain subjective experience. If subjective experience in turn is necessary to make moral judgments, computers may be limited to the realm of fact and unable to operate in the realm of morality. If legal practice and judicial practice involve moral judgments—as Part I contended—AI may be capable of playing only a limited role in the law.

Conclusion

The above analysis suggests a potential limit on AI in legal practice. It may be that AI is limited to the realm of science, that science cannot fully capture the first-person perspective, that the first-person perspective is necessary to make moral judgments, and that legal and judicial practitioners sometimes must make moral judgments. If all of that proves true, we may have identified a bulwark against AI taking over all aspects of our legal system.

To some of us, that is good news. It leaves room for human beings in the law for the indefinite future. It also provides a reason for legal practitioners—judges and lawyers alike—to acknowledge the role that moral and other value judgments play in legal practice. Many of us believe that participants in the legal process are not as transparent as they should be about how moral judgments inform the work they do.

To be sure, the above analysis does not address many practical implications of AI. Consider economics. Despite AI's limitations, it may nonetheless hold the potential to do a great deal of our legal work for us—even if people nevertheless continue to play some role on the bench and in the bar. Growth of AI in the law could be both good and bad. On one hand, if AI can improve access to legal justice, and make the legal system more affordable, that may well be desirable. Of course, the result may be more competition in the law, rendering legal practice less lucrative. But the

monopoly lawyers have on legal practice is meant to protect clients, not the financial well-being of attorneys. More generally, an animating principle of our antitrust laws is that they protect competition, not competitors. That principle would seem to apply to lawyers as well. To the extent we retain as a primary goal that competition should be allowed to flourish to the benefit of consumers, we may conclude that an expanded role for AI in law would be positive.

But our economic philosophy may need to change with technological advancements. AI threatens a disruption to our economic system as fundamental as any we have seen. If the day comes when machines can do almost all of the work we need to do, it is possible that those who control capital—who own the machines—will be able to reap the vast majority of the gains. The great masses may find themselves with very limited means to contribute to the economic system and thereby to earn a living. Never mind the special plight of attorneys. That may be lost as much of humanity flails. We may need a radical revision to how we allocate the spoils of that brave new world. Any philosophical limitations on the capacity of AI to engage in legal practice could well prove inadequate to ensure a reasonably just economic order.

This Essay does not address the looming economic and other issues that AI raises. Its scope is much more modest. It seeks merely to identify a possible limit to the role that AI can play in law (and perhaps related professions). In exploring that possible limit, it suggests that AI may advance so much that people are left with no meaningful role to perform in the scientific world—as physicists, chemists, or biologists. The same may be true for the technical tasks performed by doctors, architects, and the like. But people still may retain the unique ability to make moral judgments (and perhaps other value judgments). If so, this Essay suggests we may find ourselves in a strange new society. Imagine a young woman contemplating her major in college. She may have a passion for engineering, but she may nonetheless choose to study moral philosophy. After all, she may think, it is important to be practical, and moral philosophy is where all the remaining jobs are.