ARTIFICIAL (COMPLEX) LITIGATION

JOSHUA P. DAVIS[1]

I. Introduction

Use of artificial intelligence (AI) has exploded in recent years. Its role in society is only likely to increase. It has even begun to find its place in legal decision-making. Indeed, scholars and other commentators have speculated that computers in the not-too-distant future may displace lawyers and judges. AI may perform legal functions that traditionally have been reserved for us. This Article explores that possibility, addressing some issues that arise in particular in complex litigation.

In doing so, it focuses on two significant advances—even breakthroughs—that would likely be necessary for robolawyers and robojudges to be effective. The first would be for AI to acquire what might informally be called common sense and what more formally may be captured by the terms abductive reasoning, inference to the best explanation, and the like. This Article identifies potential challenges for programming abductive AI and speculates that we may overcome them.

A capacity for abductive reasoning might greatly expand the roles that AI can play in litigation. They could include:

(1) Strategic Advice—providing strategic advice to litigants and parties;
(2) Advocacy—advocating before a judge or jury;
(3) Class Action Settlement Assessments—assisting a judge in deciding whether to approve class action settlements; and
(4) Determining the Expected Value of Litigation—assessing the average result of a legal action, potentially as a benchmark for settlement or as a standard for imposing a result in arbitration.

A second advance—or breakthrough—would be for AI to choose its own objectives. At present, unless we provide AI ends to pursue, it is inert. This Article argues that judges in resolving cases regularly need to engage in purposive reasoning, making moral or other value judgments. That provides reason to doubt that AI will be able to fulfill the judicial role without the capacity to make value judgments. The Article suggests that AI is unlikely to develop that capacity unless and until it we imbue it with a first-person perspective—its own conscious experiences—something we have no idea how to do. So we have reason to resist the rise of robojudges.

---

[1] Professor of Law and Director, Center for Law and Ethics, University of San Francisco. Many of the issues addressed in this Article will be explored more fully in JOSHUA P. DAVIS, UNNATURAL LAW: AI, CONSCIOUSNESS, ETHICS, AND LEGAL THEORY (forthcoming Cambridge University Press 2022/23).

Part II defines some key terms and notes some of AI's recent accomplishments and failures. Part III provides a brief overview of the past, present, and possible future of AI framed in terms of four kinds of reasoning: deductive, inductive, abductive, and purposive. Part IV explores various roles AI may be able to play in litigation, especially if it improves its capacity for abductive reasoning. Part V contends that AI's inability to engage in purposive reasoning could significantly limit the efficacy of robojudges. Part VI concludes.

II. Definitions, Accomplishments, Failures

To start, we should define a couple of terms and provide a brief summary of where we are today. Consider "artificial intelligence." We will define it in a non-technical way to include all non-organic entities—which at present means computers—that can perform tasks that historically only we could do because of our cognitive capacities.[2] Examples include playing chess and Go, debating, reviewing medical images for disease, and holding conversations.

A second term that warrants definition is consciousness or phenomenal consciousness. Unless specified otherwise, we will use the word "consciousness" to mean phenomenal consciousness. Thomas Nagel famously suggested that "an organism has conscious mental states if and only if there is something that it is like to *be* that organism—something it is like *for* the organism."[3] We will use Nagel's definition, as have so many others, recognizing that it is at least conceivable that a computer could achieve consciousness.

So defined, consciousness involves first-person experiences, such as how it feels for us to see the color red, smell rotting eggs, hear Vivaldi's "The Four Seasons," taste salted caramel ice cream, or touch a worm. Each of those experiences has aspects that are directly available to us by introspection but not by other means. Only a person knows what she feels or experiences, although we can assess manifestations of those feelings and experiences and we can ask her to report on them.

With those definitions in place, let us turn to what AI has and has not achieved to date and what it may be able to accomplish. We will consider some of the current and predicted benefits and costs of AI.

AI's flashier accomplishments include beating the best human players in the world at chess and Go, prevailing over human champions at Jeopardy!, and competing remarkably well against a world class debater. AI has also shown some potential for identifying the words people read and the images they see based on their brain waves.

More mundane but also more practical is use of AI in making decisions regarding which employees to hire and promote, when to intervene to protect at-risk children in potentially unsafe homes, what the terms of bail should be for criminal defendants in light of the odds of flight or recidivism, how to target advertisements to maximize their efficacy, and whether to take

---

[2] We might broaden the definition to include tasks that intelligent non-human animals can perform because of their cognitive capacities. Nothing essential to the analysis should turn on this definitional issue.

[3] Thomas Nagel, 83 Phil. Rev. 435, 436 (1974).

remedial actions for possible cancer and other diseases based on medical images. AI may soon drive our cars, do our shopping, and engage in routine daily communications for us. It has already made substantial progress in each of those tasks and many others.

Some of AI's current and future accomplishments are and will be highly beneficial, even if they may come with costs. In looking for cancer, for example, AI can enable us to decrease false negatives—and delayed treatment—as well as false positives—and unnecessary surgeries. We should welcome such progress, notwithstanding that AI may put some radiologists and other medical services providers out of work.

Similar points apply to legal practice. Relying on AI do some legal work—undertaking document review, for example, or drafting simple contracts—can help us meet some of the desperate need for modestly priced legal services. Again, that is an improvement worth exploiting, even if it eliminates some human jobs.

We might say the same about autonomous vehicles. The lives they can save and the injuries they can avoid, and the hours they can free up for workers during commutes, should prove valuable. That is true even though drivers for taxi and ride-sharing services would suffer as a result.

As with displaced doctors and lawyers, such job losses should be taken seriously. Still, they would not be unique. Technological change has always displaced workers. No one today is employed by buggy whip factories. Even if AI causes disruption to the work force—as some predict—there should be ways to distribute society's gains broadly, perhaps through a universal basic income. Addressing that issue, however, is beyond the scope of this Article.[4]

We will be concerned, however, with whether AI would do more harm than good in other ways. Consider the potential for racial and other biases in employment decisions, juvenile dependency proceedings, bail, and advertising. In each setting, excitement about AI has given way to concerns that it embodies our biases. In some cases, it is a cruel irony that AI may exacerbate the very forms of discrimination it is intended to ameliorate.

Similarly, consider the role of AI in guiding Internet users. Stuart Russell, an eminent AI expert, explains that content-selection algorithms on social media are often "designed to maximize *click-through*, that is, the probability that the user clicks on presented items."[5] Presumably, the algorithms were expected to select internet links to suit users. But that is not all they did. They molded user preferences, enabling more reliable predictions of which links the users would click. Russell suggests that this phenomenon can explain why social media tend to direct users to Websites that foster extreme political views. Political extremists have more predictable preferences than do moderates.[6]

---

[4] Note controversies about whether on net AI will decrease or increase employment and about how new jobs will be distributed.
[5] Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control 8 (2019).
[6] *Id*. at 8-9.

With the above sketch of some of the potential and perils of AI in mind, we should consider two of its limitations. The first involves instrumental reasoning. It currently struggles in pursuing certain goals we assign it because it lacks what we might colloquially call common sense. We cannot ask it to avoid harmful discrimination or the inculcation of hate. At present, we would have to define those and similar goals in much more technical and precise terms. AI is incapable on its own of formulating hypotheses about the kinds of links that might encourage invidious discrimination or hate.

A second limitation relates to what we will call purposive reasoning. Current AI cannot decide for itself what purposes to pursue. It cannot form its own objectives. We must choose them for it. Unless so directed, AI would not on its own decide not to promote bias or hate.

The next section places these points in historical context. In doing so, it distinguishes four forms of reasoning, describes efforts to program AI to engage in those forms of reasoning, and suggests some legal tasks those forms of reasoning may enable AI to perform.

III. Four Phases and Four Forms of Reasoning

We can understand the progression of AI in terms of four phases, each defined by a form of reasoning: deductive; inductive; abductive; and purposive.

To oversimplify, early AI was largely deductive. We supplied AI with general rules and it applied them. In recent decades—with the rise of so-called Big Data—the emphasis has shifted to induction. We provide AI general goals and a framework for engaging in inductive reasoning, particularly in the form of statistical analysis. AI then detects patterns, enabling it to derive its own rules. To be sure, AI applies those rules deductively. But much of AI's increased power comes from inductive reasoning—from distinguishing patterns in data from noise.

Two additional phases have not yet arrived and may never do so: abductive AI and purposive AI. Abductive reasoning involves what one might call common sense, especially as it is used to formulate testable and working hypotheses. Plausible explanations for how the world works can inform what propositions are worth testing and what assumptions or premises we should adopt until we encounter evidence that contradicts them. At present, AI arguably does not engage in abductive reasoning (or does so only to a limited extent). We have to do abductive reasoning on AI's behalf, relying on our own hypotheses in programming deductive and inductive AI. One might say that AI lacks common sense.

Purposive reasoning involves forming objectives. Deductive, inductive, and abductive reasoning are instrumental. They provide means of achieving prescribed goals. We are capable of forming objectives. We have preferences, aversions, values, and the like. AI does not. It is inert without our guidance. We have to tell it what ends to pursue.

Prognostication is notoriously hazardous, particularly about AI. That said, Part III suggests that AI may well improve significantly at abductive reasoning but that it is unlikely to become capable of purposive reasoning. Put more simply, Part III predicts that AI will acquire some common sense but it will not be able to choose its own ends. The rest of the Article then assumes the accuracy of those predictions.

A. Deductive Reasoning: Early AI

For many decades, AI employed deductive reasoning. We created algorithms directing it how to respond to input. AI would then apply those algorithms in a mechanical fashion. Some of what deductive AI could accomplish was and is impressive. It can solve mathematical problems that would take most of us a great deal of time and effort, if we could handle them at all.

To be sure, whether a calculator should qualify as AI could be fairly debated. Precisely because it is purely deductive, one might argue, it does not require judgment and so lacks a quality many associate with intelligence. As defined in Part II, however, intelligence includes everything that historically only human beings could do because of our cognitive capacities. That includes solving many mathematical problems involving non-trivial addition, subtraction,[7] multiplication, division, exponents, and the like. So, for our purposes, we will say that a calculator is a form of AI.

One of the simplest forms of deduction involves a syllogism. To illustrate the different forms of reasoning we will be discussing, it will be helpful to use variations on a concrete example. Here is one for deduction:

Example 1: Deductive Reasoning

Proposition 1: All fish swim.

Proposition 2: Tuna is a fish.

Conclusion: Tuna swim.

An advantage of deduction is that it can be unerring. If a deductive argument is valid and sound, the conclusion that follows is correct. A deductive argument is valid if the conclusion must be true given that the premises are. The argument is sound if its premises are true. In other words, if all fish swim, and if tuna is a fish, then tuna necessarily swim. That deductive argument is valid. It also would be sound if all fish swim—which is at least pretty close to true. To be sure, some injured fish do not swim. Also, for example, the red-lipped batfish found in the waters off the Galapagos Islands uses its pectoral fins to walk or, rather, to stagger like a drunk (and it is shaped like a bat and has bloated red lips and a unicorn horn to boot).[8] Still, the vast majority of fish swim. So the above syllogism generally holds true, even though it is not infallible in the way that deduction generally should be.

Even if deductive arguments hold the potential for reliability, they are limited in important ways. One of them is that they merely build on or extend what we already know. We have to supply the premises, such as that all fish swim. Deduction does not allow us to infer that all fish swim, unless we can derive that conclusion from more general rules.

Still, deductive AI has had some interesting and valuable applications. They include computers capable of playing chess or Go better than many of us can. For years, advances in the

[7] Note some animals can arguably do simple addition and subtraction.
[8] https://en.wikipedia.org/wiki/Red-lipped_batfish

quality of AI chess and Go resulted in part from improvements in algorithms we created and the speed and memory of computers. We might program a computer to understand the rules of chess. We then might instruct it to give different weights to different chess pieces—say, a queen ten points and a pawn one point—and to different positions on the board—say, a rook on the seventh rank or to two rooks on the same rank or file. The computer could then be directed to assign points to various available moves by analyzing all of the possible responses to them, responses to the responses, etc.

There are too many permutations in chess—or Go—for deductive AI to assess them all. So the inquiry would be limited by time and memory. Still, deductive AI made substantial progress in this way. But it never beat the best human beings in the world at chess or Go. That had to await development of *inductive* AI.

We see important parallels in legal work. Word processing is an example. Word processing programs—like Microsoft Word—relied historically on deductive reasoning. We press a key or a number of keys and our computer applies a rule for what actions to take. That technology has had a profound impact on legal practice. Large numbers of typists lost their jobs. Many more lawyers today than in the past do their own word processing. And the ability of attorneys to share and modify documents via computers has transformed their lives.

Other legal tasks, even seemingly menial ones, are not susceptible to deductive reasoning. Consider document review. It is hard to anticipate the characteristics of a document that make it important to a case or that render it protected by the attorney-client privilege or work product doctrine. General rules are clumsy for those purposes. Deductive AI thus had only limited utility at reviewing documents in place of (junior) attorneys.[9] That changed with improvements in inductive AI.

B. Inductive Reasoning: Current AI

Recent decades, including the era of Big Data, have seen the rise of inductive AI. Induction involves learning from experiences or observations. Over time, we discern a pattern and expect it to repeat in the future.[10] A set of formal rules we have developed to test for patterns is known as statistics (or sometimes econometrics). Inductive AI could be described as statistics on steroids.

Inductive AI has made extraordinary strides for several reasons. We have become much more sophisticated at programming AI to perform statistical analyses. The speed and memory of computers have increased dramatically. So has access to data, particularly from the Internet. That last factor is why we use the term Big Data, but we might do better to call it the age of Inductive AI instead.

Regardless of labels, inductive AI performs many impressive tasks. We will discuss some of them below, but first consider a simple example of how inductive AI works. We might

[9] Cite Rick Marcus; other sources.

[10] Acknowledge Hume's famous critique of the philosophical basis for induction and ongoing disagreement about whether it has been addressed adequately.

provide inductive AI with a very large number of observations of fish and enable it to determine whether they are swimming. One way to do that would be to develop a large dataset of videos of fish (and perhaps other creatures and objects), some of which are swimming and some of which are not. We would label each observation as "swimming" or "not swimming." Through iterated rounds, we could then train the inductive AI to assess whether each video contains a fish that is swimming. It could become accurate at doing so. We could then apply the inductive AI to videos of fish to see whether they all swim. It might conclude that they all do (or that a very high percentage of them do). If we inform the inductive AI that tuna is a fish (or similarly teach it to recognize fish), it could then combine inductive reasoning with deductive reasoning to perform the following analysis:

Example 2: Inductive Reasoning

Proposition 1: The fish we observe (virtually) all swim.

Proposition 2: Tuna is a fish.

Conclusion: Tuna (almost certainly) swim.

An advantage of inductive AI is that it can form its own rules. It can identify patterns of which we are not aware. An example is an algorithm that analyzed images on dating sites and was able to predict sexual orientation with a high rate of success. The programmer did not build the algorithm to perform that task. Nor did he know how the inductive AI worked—what features it used to make its determinations. He merely tasked his algorithm with finding patterns in images in various ways and—lo and behold—it was reliable in particular at distinguishing homosexual men from heterosexual men from their photographs.[11]

Inductive AI has various limitations. One of them is that it offers probabilistic predictions, not certainy. Induction is not unerring, the way deductive reasoning can be. At best, it offers reliable statements about the likelihood of different possible outcomes.

Further, although inductive AI is, in the above sense, not as dependent on human input as deductive AI, we still need to play a significant role in programming it. We have to identify the variables it will use for analysis—such as pixels in videos—although inductive AI may then group those pixels depending on patterns it discovers. We also have to label the data to match the variables. And we have to build the framework for the statistical analysis. We put in place what are sometimes called its "hyperparameters."[12]

Inductive AI also has a drawback: it may embody biases or draw inferences in ways we find inappropriate or offensive. To build on the example above, once inductive AI can predict the sexual orientation of men with accuracy, it may then use that knowledge to discern other patterns

[11] Cite original incident. See also https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph (later more successful effort in 2017 to identify sexual orientation from photographs).
[12] See Melanie Mitchell, Artificial Intelligence: A Guide for Thinking Humans 97-98 (2019).

and make predictions. Those could be objectionable. They could also reflect outcomes that are shaped by societal prejudices.

Relatedly, inductive AI would not have the common sense to worry about predictions based on sexual orientation or to report them as part of its results (unless we instruct it do so). We thus may not know—or even have the capacity to know—that AI has discerned and is relying on patterns based on sexual orientation, race, sex, religion, or other similar categories. Nor would AI be capable of making value judgments that might steer it away from what may be stereotypes.

Consider the problems that can arise if inductive AI uses data that is tainted by biases. It may use raises and promotions, for example, to assess worker potential. Those raises and promotions may reflect sexism, racism, homophobia, or religious intolerance. If so, then inductive AI will make predictions that embody those biases, possibly perpetuating invidious discrimination. As statisticians like to say: garbage in, garbage out.

Subject to the above limitations and concerns, inductive AI has accomplished a great deal. It is used for the many remarkable successes noted in Part II in chess, Go, Jeopardy!, and debate, as well as in spotting cancer in medical images, interpreting brain waves, assisting employment decisions, identifying at-risk children in homes, predicting recidivism, and targeting advertisements. It also poses the dangers we have recognized, including exacerbating stereotypes and inequalities.

Both dynamics are reflected, for example, in use of facial recognition software to ensure that unemployment benefits go to appropriate recipients. Inductive AI has the potential to be far more efficient, cost-effective, and accurate than we are at that effort. It also has fared much worse for dark-skinned people than light-skinned people and much better for people whose gender expression remains stable over time than for those whose gender expression changes.[13] That can create troubling disparities in access to government funds.

Inductive AI also has been responsible for significant improvements in automating litigation. Think again about document review. In III.A, we noted that deductive AI has a limited capacity to review documents for relevance or privilege. We struggle to fashion general rules for deductive AI.

Inductive AI can be much more effective. We can provide inductive AI a set of sample documents, some of which we label as "hot" or "protected" and others we do not. AI can then use statistics to identify characteristics of the documents that are indicative of relevance or privilege. After iterations, inductive AI can then be set loose on a huge volume of documents,

[13] See, e.g., https://www.technologyreview.com/2021/09/28/1036279/pandemic-unemployment-government-face-recognition/?truid=&utm_source=the_download&utm_medium=email&utm_campaign=the_download.unpaid.engagement&utm_term=&utm_content=09-28-2021&mc_cid=9c574f15ea&mc_eid=8cec2fb502 (last visited 9/28/2021)

separating relevant from irrelevant documents and protected from unprotected ones with great speed, efficiency, and reliability.

But note the active role we have to play in the process. Inductive AI is not capable on its own of making reasonable *preliminary* or *working* judgments about which documents are apt to be relevant or privileged. It lacks common sense, situation sense or reasonable judgment to perform that task. As discussed next, that would seem to require abduction.

C. Abductive Reasoning: The Next Breakthrough?

The next category of reasoning is more difficult to define in part because we do not understand it well. A formal word for it dates back to the American philosopher Charles Sanders Peirce: abduction.[14] We have already described it as common sense, although that term is much broader and looser. We might say that abduction involves one or more categories of common sense.

The first such category comprises testable hypotheses. That was Peirce's focus. In discussing abduction, he was concerned with how we initiate the scientific method by generating hypotheses worth testing. To take a concrete example, we might see smoke coming from the windows of a house. That could lead us to form the testable hypothesis that the house is on fire. (The relationship to common sense should be clear.) We might then follow up with an immediate investigation, peaking inside to look for flames.

A second category of abduction comprises working hypotheses. We accept certain propositions as true unless and until we have reason to conclude otherwise. If we see black smoke billowing from a house, we may not wait to investigate before calling the fire department. We might feel confident enough to act right away. Our working hypothesis is that the house is burning. We act on that hypothesis until we have a firm basis to conclude it is wrong.

We will use the term abduction to include forming both testable and working hypotheses. Admittedly, they may be related or distinct—and it could be that each should be divided into multiple forms of reasoning. Part of the problem is that we have not reached a consensus either in theory or in practice about abduction.

Some philosophers consider abduction to be a form of what they call inference to the best explanation.[15] Inference to the best explanation entails, as the name suggests, drawing the inference that would provide to the most compelling explanation for a state of affairs. To continue the above example, on seeing smoke coming from a house, we might conclude that it is most consistent with our understanding of the world that fire would be the cause. Fire may offer a better explanation than, say, a smoke machine or magic. That would be true for various reasons. Smoke machines are relatively rare. Magic conflicts with our understanding of the laws

[14] See, e.g., Erik J. Larson, The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do 25-26, 99-102, 160-68, 190 (2021); Douglas Walton, Abductive Reasoning 3-17 (2005).

[15] For an excellent discussion of the topic see Peter Lipton, Inference to the Best Explanation (2d ed. 2004).

of science. At least in terms of probability and conformity with our other beliefs, fire may provide the best explanation for the smoke.

To be clear, what counts as the "best" explanation of a phenomenon is far from straightforward, a complexity we will not explore.[16] Relatedly, abduction also may or may not be a form of inductive reasoning. But if it is, it is one we have been in significant part unable to get AI to do on its own.

Before developing that point, however, let's return to variations on the example we used for deductive and inductive reasoning. They will help us to see abduction's potential and its limitations. Here's an example:

Example 3: Abductive Reasoning

Proposition 1: Fish swim.

Proposition 2: Tuna swim.

Conclusion: Tuna is a fish.

Abduction can help us identify the possibility that tuna is a fish from our knowledge that fish swim and that tuna do too. That conclusion does not follow deductively. While it may be true (let's assume) that all fish swim, it is not true that everything that swims is a fish. Nor may we have a basis in inductive reasoning to infer that an object that swims is a fish, perhaps because we have not studied what percentage of swimmers are fish. So there seems to be an element of informal or unsystematic judgment arriving at the testable or working hypothesis that tuna is a fish from the observation that it swims.

Consider another example:

Example 4: Abductive Reasoning

Proposition 1: Fish swim.

Proposition 2: You swim.

Conclusion: You are a fish.

As you can see, abduction can easily go awry. It does not incorporate formal standards for reliability like the ones that we have developed for deduction and induction. Still, *testable* hypotheses play a key part in advancing knowledge. That is true even for propositions that may at first seem absurd—such as that time moves at a relative rate or that electrons do not occupy a single position at any given time or that light is both a wave and a particle.

Further, intuitively plausible *working* hypotheses are often essential for action that is timely and effective. You may not have seen a house fire before—and so you may have no idea

[16] For a sophisticated discussion of the relevant meaning of explanation see id. at 21-54.

of its odds of occurring—but you still likely should call for help right away if you see smoke billowing from someone's home.

Note along these lines that the difference between example 3 and 4 is obvious to us but not necessarily to AI. We have to program AI to detect absurdities. That is something we have struggled to do. It is no mean feat to anticipate the messiness of reality such that abductive AI can function effectively.

As a result, we have a capacity to navigate our everyday lives that AI lacks. AI may beat us at chess and Go, but we are superior to it at walking down the street, making appropriate small talk, and generally making sense of our physical and cultural environments. Melanie Mitchell aptly encapsulates this point: for AI hard things are easy and easy things are hard.[17]

We do not know enough about our own abductive reasoning—or common sense or inferences to the best explanation—to operationalize a similar capacity in AI. But we can speculate about why we have advantages over AI at abductive reasoning (speculation that is itself abductive). For example, evolution may have caused us to develop particularly effective default structures or innate schemas for making sense of the world. Technologists have found some support for this view when it comes to inductive reasoning. At times they have made progress by mimicking our biological structure in setting the hyperparamaters for inductive AI, often without knowing why our neurological systems work as well as they do.[18]

Perhaps we are hard-wired to recognize particular patterns in the world. Newborn infants, for example, respond differently to human faces than to other objects. They may have some built in capacity for face recognition.[19] Similarly, many scholars believe that children learn languages far more effectively than can be explained merely by induction from their observations of others speaking. One theory is that children have an extraordinary knack at discerning patterns of communication. Noam Chomsky famously argued, for example, that infants are born with a framework for understanding grammar.[20] Even those skeptical of Chomsky's views tend to accept that human brains predispose us to certain aspects of language acquisition, if based only on our natural capacity to learn in general.[21]

---

[17] Melanie Mitchell, Artificial Intelligence: A Guide for Thinking Humans 33 (2019).
[18] *Id*.
[19] *See, e.g*., Mark Johnson, Face Perception: A Developmental Perspective in Oxford Handbook of Face Perception (Nov. 2012) (G. Rhodes, A. Calder, M. Johnson, & J. Haxby, eds.) (available at https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199559053.001.0001/oxfordhb-9780199559053-e-001).
[20] *See, e.g*., Noam Chomsky, "Approaching UG from Below," in Hans-Martin Gartner & Uli Sauerland, eds, Interfaces + Recursion = Language? Chomsky's Minimalism and the View from Syntax-Semantics (2007); Noam Chomsky, A Review of B.F. Skinner's Verbal Behavior, 35 Language 26 (1957).
[21] *See, e.g*., Vyvyan Evans, The Language Myth (2014); Geoffrey Sampson, The "Language Instinct" Debate (rev'd ed. 2005).

We are far from simulating human thought through abductive AI. Yet it seems plausible that we will make progress on that front. It may be that we need to develop further knowledge about how our neurological systems are structured and what assumptions or schemas we have in place to facilitate learning. We might then be able to program abductive AI using the same structures and assumptions or schemas that we do.

Perhaps abductive AI could simulate our thinking in the near term and then improve on it in the longer term. In a sense, the pattern might echo the progression from deductive AI to inductive AI. At first, abductive AI may rely on substantial direction from us—top down. Over time, abductive AI may be able to detect its own patterns, including ones we miss—bottom up.

There is a potential analogy to—or even mapping onto—what Daniel Kahneman calls fast thinking and slow thinking.[22] Fast thinking is quick, easy, automatic, intuitive, and subject to various cognitive biases. Slowing thinking is plodding, difficult, deliberate, formal, and more resistant to cognitive biases. Abductive AI might progress at first by simulating our fast thinking, cognitive biases and all. It may then shift more and more to slow thinking—albeit performed quickly, possibly far more quickly than we could.

To be sure, fast thinking may never disappear entirely. We may understand fast thinking as involving, at least in part, pre-set theories about the world. There is a credible argument that we cannot identify any facts about the world—make any observations—without some implicit theory. As Hilary Putnam quotes A. E. Singer, "Knowledge of (particular) facts presupposes knowledge of theories (that is, of generalization)."[23] Still, abductive AI may be able to pare back our least reliable implicit theories, holding on to only a bare minimum that allows it to frame inductive inquiries and thereby detect patterns that elude us.

Regardless of the precise route, we will assume for the rest of this Article that abductive AI proves tractable. AI has arguably made some progress at abduction, depending on how that term is defined. Although AI often struggles with the real world, the situation is improving. Self-driving cars are able to navigate real roads. They have experienced some catastrophic failures, and many glitches, but they also succeed with promising regularity.

More generally, to the extent that abduction merely involves instrumental reasoning, and to the extent instrumental reasoning is distinct from choosing objectives, AI may eventually engage in abductive reasoning as well as us or even far better than us. That may take a while. It did for chess and Go. Indeed, some commentators famously predicted that chess at the highest level requires intuition of a sort that computers lack. A human being, they claimed, would always be the world champion. Those predictions and their justifications proved mistaken. Whatever intuition is involved in abduction may turn out to be similarly conquerable by AI.

If it is, abductive AI may be able to engage in all sorts of instrumental tasks that at present only we can do. One of them could be making preliminary or working judgments about which documents are likely to be relevant to a case and about which ones are likely to be

[22] Daniel Kahneman, Thinking, Fast and Slow ().
[23] Hilary Putnam, The Collapse of the Fact/Value Dichotomy and other Essays 141 (2002).

protected from discovery. Abductive AI might then be able to carry much more of the burden of document review than can inductive AI. It might not need as active and extensive human supervision.

Abductive AI also might be able to make the common-sense judgments necessary for effective advocacy, particularly oral argument. It might formulate effective working hypotheses about the kinds of assertions and interpretations that judges or jurors would find persuasive and credible. Even in the absence of a systematic empirical analysis—and perhaps even if one is impossible or infeasible—abductive AI might be able to form hunches and make educated guesses. Those could be crucial for it to prove persuasive, especially in the real-world environment of a courtroom.

Note, however, the important qualification above about the difference between instrumental reasoning and choosing objectives. It could be that some of what we call abductive reasoning—or common sense or inference to the best explanation—requires more than instrumental reasoning. It may also depend on a sensible selection of purposes or goals.

Here we come to a central debate in the philosophy of science. Hilary Putnam, for example, argued that science depends on "epistemic" values, [24] that is, values we pursue to help us make sense of the world. They include coherence, plausibility, reasonableness, simplicity, elegance, and beauty.[25] Those values may play an essential role in our selection of testable and working hypotheses, including what we call hunches and educated guesses. Our skill at abductive reasoning may rely in part on our assessments of whether a theory is elegant or beautiful. Yet we may not be able to program AI to make those kinds of assessments. Doing so may require a choice among objectives—or giving content to selected objectives, which may amount to the same thing. As we discuss next, AI may be incapable of that task.

D. Purposive Reasoning: Future or Fantasy?

Where AI faces its greatest challenge is in purposive reasoning. That is the term we will use for the capacity to choose objectives. Unlike deduction, induction, and even possibly some aspects of abduction, programmers have made no progress in getting AI to select its own ends. Stuart Russell, one of the world's leading experts on designing AI,[26] explains, "Because machines, unlike humans, have no objectives of their own, we give them objectives to achieve. In other words, we build optimizing machines, we feed objectives into them, and off they go."[27] We need to give AI its objectives. It cannot form them on its own. For now, that is a brute fact.

---

[24] Hilary Putnam, The Collapse of the Fact/Value Dichtomy and Other Essays, at 30-34,132, 135, and 143. For a fascinating discussion of the role of beauty in science in general and physics in particular see Steven Weinberg, Dreams of a Final Theory: The Scientist's Search for the Ultimate Law of Nature 90-165 (1992).

[25] Id. at 132, 135, and 141.

[26] See, e.g., Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach (4th ed. 2021).

[27] Human Compatible at 10.

The same seems to be true for AI's lack of consciousness. As far as we know, it does not have any. Nor do programmers have any idea how to imbue AI with consciousness. Russell again: "In the area of *consciousness*, we really do know nothing. No one in AI is working on making machines conscious, nor would anyone know where to start, and no behavior has consciousness as a prerequisite."[28]

These points suggest two important possibilities. The first is that there may be a relationship between AI's apparent lack of consciousness and its inability to form objectives. In particular, AI's lack of consciousness may *explain* its inability to form objectives. Further, AI's inability to form objectives may be important *evidence* of its lack of consciousness. A second—and related—possibility is that Russell may be wrong that consciousness is not a prerequisite for any behavior; it may be essential for forming objectives.[29]

Let us begin with why phenomenal consciousness might be necessary to form objectives. At least since the Scientific Revolution, we have proceeded based on the assumption that only conscious beings are purposive. We have put aside Aristotelian and other forms of teleology as a way to describe mindless nature. We do not predict the effects of gravity by describing the strivings of earth, water, air, and fire to sort themselves out in layers, from the bottom to the top, in the order listed. Instead, we use the laws of Newtonian physics or, at the extremes of scale, quantum mechanics and relativity. The laws of the physical sciences are causal, not purposive.

Current AI operates in the realm of science. It follows formal rules of causation. Those rules may be complicated. They are, however, ultimately deterministic—or, on a very small scale, probabilistic. We set the train of AI in motion in the direction we choose. It optimizes what we instruct it to optimize.

Human beings, in contrast, are motivated by desires, aversions, aspirations, and at least arguably values (more on that below). Science can help us describe what is, predict what will be, and help us to manipulate what is so that we bring about what we wish to be. We alone have wishes. Stars, planets, gravitation, and photons do not.[30]

---

[28] Human Compatible at 16 (emphasis in original).

[29] In exploring these possibilities, we will skirt some deep philosophical issues by noting an important distinction between theory and practice. The analysis that follows focuses on what is feasible in the foreseeable future. It does not seek to resolve profound disputes that have vexed philosophers for centuries and, in some cases, for millennia.

[30] To be sure, some schools of thought have attempted to eliminate the mind and notions of purpose in describing, predicting, and manipulating human action. Notable in this regard were the radical behaviorists. Although the field is still alive, it has fallen far short of its original ambitions. Similarly, some philosophers claim that science—first and foremost physics, but also chemistry, biology, and some branches of psychology—will displace talk of the mental. Dennett; Humphrey. That possibility should be taken seriously. Philosophical materialism may well prove correct. We may someday reduce our phenomenal experiences to physical processes—or at least establish that our minds have no causal force and are merely epiphenomenal. But, for now, the best way to explain human behavior is in part in terms of motivations, that is, as purposive.

Let's consider an example to be clear what we mean by purposive reasoning.

Example 5: Purposive Reasoning

Proposition 1: We should (or should not) eat fish.

Proposition 2: Tuna is a fish.

Conclusion: We should (or should not) eat tuna.

Here Proposition 1 contains a value judgment. It says what we should (or should not) do. The value judgments it makes could be justified in various ways. Perhaps we *should* eat fish because doing so is healthy for us, as well as more humane and better for the planet than eating cows and pigs. On the other hand, perhaps we should *not* eat fish because they are killed or farmed inhumanely, because it is wrong to destroy a sentient life unnecessarily, or because it is bad for the planet. Those and other possibilities may be justified ultimately on various grounds: for example, deontological, consequentialist, religious, or cultural.

Regardless, the point is that AI cannot make value judgments on its own, as Russell acknowledges. Moreover, as Russell contends—it is a theme of his book—we do not know how to program AI with sufficiently general, adaptable, and reliable objectives for us to trust it to operate independently. His solution is to build AI so that it seeks guidance from us about our preferences on an ongoing basis. That is a possible strategy for contending with the risks posed by AI—one that should be taken seriously even if, as I contend elsewhere, it has significant drawbacks.[31]

Particularly relevant for present purposes, Russell's ultimate position—that we should force AI to consult with us regularly—is in tension with his claim that consciousness is not a prerequisite for *any* behavior. It may be a prerequisite for *forming objectives*. That behavior—as Russell notes—is essential. Hence his belief in the need for human beings to steer AI.

We have assumed that we will figure out how to build abductive AI. We will not make the same assumption about purposive AI. As noted above, Russell acknowledges that technologists do not have any idea how to imbue AI with consciousness—nor any idea where to start. As he reports, no one is even working on that project. So we will assume that AI will continue to lack consciousness and, as a result, it will remain incapable of purposive reasoning. We turn next to the implications of these assumptions for the future of AI in litigation.

IV. Applications to (Complex) Litigation: Prediction and Manipulation

If AI continues to improve its instrumental reasoning, it may become a powerful legal tool. We should expect it to make accurate predictions about litigation, identifying potential outcomes and their odds of occurring. After all, AI can detect patterns. That is why inductive AI is like statistics on steroids. In effect, it performs facial recognition by predicting which combinations of pixels in an image will be associated with a specific person, it reads minds by

[31] See Joshua P. Davis, Unnatural Law: AI, Consciousness, Ethics, and Legal Theory (Cambridge University Press 2022/23).

predicting which brain waves will be associated with which words or images, and it wins at chess and Go by predicting which moves will be associated with the highest probability of winning a game. As Agrawal, Gans, and Goldfarb point out, inductive AI's core strength is as a prediction machine.[32]

Abductive reasoning should fortify that strength. It could enable AI to formulate testable hypotheses that we miss, just as inductive AI currently draws inferences that we miss. Abductive AI might also develop effective working hypotheses that help it anticipate how judges and juries would respond to the law, evidence, and other cues about the merits of a case.

Nor should we overlook that prediction can be tantamount to manipulation. If we can anticipate how human beings will respond to stimuli, we may be able to shape the environment to elicit the behaviors we want, including from judges and jurors.

As discussed next, AI that can predict and manipulate could have great utility. Tasks it might perform include providing strategic advice, advocating for clients, helping judges assess proposed class action settlements, and proposing or imposing attractive compromise outcomes.

A. Strategic Advice

One use of AI could be to help litigants and lawyers act strategically. They could consult AI, for example, in deciding whether to settle a case and, if so, on what terms. AI could also guide their conduct in litigation—in electing which witnesses to call, what legal and factual arguments to make, and what evidence to introduce.

At present, attorneys often act on hunches at worst and on experience at best. But experience without reliable feedback does not yield expertise.[33] Lawyers may think they know what persuades judges and juries. That does not mean that they do. With judges, at least, they may receive some regular feedback—immediate or delayed—as a guide. But it is hard to know what actually drives judicial behavior. And few lawyers appear before juries frequently enough and get feedback from them regularly enough to have a sound basis for predicting what evidence and arguments will prove persuasive. Nonetheless, evidence suggests that experienced practitioners in a highly-skilled field, buttressed by a strong professional culture, will tend to have an exaggerated sense of their capacity to make sound predictions.[34]

AI may be able to provide a much more reliable, empirical basis for strategic decisions. It could conceivably obtain information from real and mock oral arguments and trials. Using that data, AI might process images, sounds, words, and movements, and associate them with successful and unsuccessful advocacy.[35] Based on those empirical analyses, it might provide

[32] Ajay Agrawal, Joshua Gans, and Avi Goldfarb, Prediction Machines: The Simple Economics of Artificial Intelligence (2018).
[33] See Kahneman, Thinking, Fast and Slow 234-44 ().
[34] Id. at 209-21.
[35] The data might come from observations that are familiar to us—such as cameras that capture light we can see and microphones that detect noises we can hear—as well as ones that are not—

valuable guidance on what arguments attorneys should make, what phrases they should or should not use, how to frame witness testimony, which lawyers are most likely to influence the relevant decision-makers, what clothes and accessories attorneys and witnesses should wear, what body movements they should make, how they should adjust their facial expressions, and a host of considerations that might not occur to most of us but that may matter in practice.

AI also may be able to tailor those recommendations to particular judges[36] and jurors. It might do so based on observable facts about them—their appearances, their occupations, the places they live, or their accents. It also could potentially rely on a treasure trove of information gathered about most Americans based on their online activities. Litigators could potentially feed AI the data that is collected about what products individual decision-makers buy, what services they purchase, what internet sites they visit, what emails and texts they send, what documents they create, what property they own, what friends they have, where they travel throughout the day as recorded by their cell phones, and anything else that is collected and packaged about us from the Internet and that AI determines is relevant. After all, we live in what has been called the Age of Surveillance Capitalism.[37] All of this data might give AI insights into how to manipulate judges and jurors in ways we would not predict and may not even be able to understand.

AI also should be better able to assess its own limitations than can human experts. Again, we tend to be more confident than accurate.[38] As Oliver Wendell Holmes, Jr. observed, "Certitude is not the test of certainty. We have been cock-sure of many things that were not true."[39] AI should better than we are at making predictions and at acknowledging uncertainty. Both can be valuable.

B. Advocacy

AI may be able to do more than just provide lawyers and clients strategic advice. It may be able to advocate. AI may someday synthesize legal sources, evidence, allegations, or the like in a written form that maximizes the chances of bringing about a desired outcome. It may do so by appealing not only to a judge's values but also to her biases and confusions. In this regard, too, AI may exploit information about a judge, perhaps from public records, including past judicial proceedings, and perhaps from her online activities, much of which is or may be for sale.

In the not too distant future, we also may be able to build AI that looks and sounds like a particularly credible advocate. Perhaps AI may be able to build such AI.[40] And, again, the look

---

perhaps involving hyperspectral light, high-frequency sounds, thermal energy, and who knows what else.

[36] Note the French law that bans this use of AI.

[37] See Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (2019).

[38] See Kahneman, Thinking, Fast and Slow 209-221 ().

[39] Oliver Wendell Holmes, Jr., Collected Legal Papers 311 (1920).

[40] Note predictions about the singularity and Larson's skepticism.

and sound of a roboadvocate could even be bespoke—tailored to a particular judge or set of jurors.

To be sure, those sorts of achievements would not come easily. Technologists have struggled with natural language. We have struggled to make AI that replies to written or spoken words in ways that human beings find easy.[41] It is possible that doing so would require AI to move beyond syntax to semantics[42]—not only identifying patterns in words but understanding what those words mean.[43]

On the other hand, it is also possible that AI will become so sophisticated, fast, and powerful that it will be able to write more effective briefs than we can without any understanding of what they mean. It may simply predict our responses to letters, words, sentences, or paragraphs, much as it associates pixels in an image with a person's identity. Let's assume that it does.

We still may retain an advantage over AI in advocacy. Although in theory attorneys do not testify when they advocate, seasoned observers suggest that credibility is essential to persuasion.[44] We have assumed AI is not conscious and so it cannot be sincere. That might undermine a judge's confidence in AI arguments.

That problem might not be as significant for AI briefs. It is not clear that a judge ever needs to know that a machine was the author, at least as long as an attorney signs it (and, presumably, reviews it before filing). But it would be more difficult to mask that AI is arguing in court. So AI might operate at a net disadvantage in oral advocacy.

But, then again, it might not. We know that people form strong relationships with inanimate objects, even feeling love for them.[45] We have a powerful propensity to anthropomorphize. A well-designed robot might take advantage of that propensity. It might even be tailored so we find it particularly credible. It could rely on information about a decisionmaker's background, facial expressions, accent, and the like to adjust its gestures, posture, voice, and appearance. Presumably judges should not be influenced by these factors. Yet they likely are. So robolawyers may someday be more effective than human lawyers in all forms of advocacy.

C. Settlement Approval in Class Actions

[41] See, e.g., Erick J. Larson, The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do 50-59 (2021).
[42] Cite Searle.
[43] Larson; Mitchell.
[44] See, e.g., Herbert J. Stern, Trying Cases to Win.
[45] See, e.g., https://www.thedailybeast.com/why-humans-love-robots-like-people

Another potential use of AI would entail assisting judges in a difficult task in class actions. They have an obligation to protect absent class members by assessing whether a proposed class action settlement is fair, reasonable, and adequate.[46] That is no mean feat.

When class litigation turns to settlement, the adversarial system breaks down. In litigation, plaintiffs and defendants have incentive to offer competing views of the law and facts. That assists the judges in reaching informed conclusions.[47] When the parties settle, however, they present a unified front. The judge is largely on her own in determining whether plaintiffs have obtained sufficient relief for the class.

True, objectors may challenge a class action settlement. But they often have a limited capacity to assess the relevant law and evidence—and a limited interest in doing so.[48] Indeed, they often seek merely to gum up the works until they are paid to go away. So a judge may find herself in the unenviable position of second-guessing attorneys who know the law and the facts far better than she can.

Enter AI. A judge could use its analysis of the likely outcomes of litigation and their odds of occurring. That could assist a judge in assessing the relationship between a settlement and what might be expected to happen on average in litigation. AI could also identify any extreme results that might occur, a consideration relevant to whether a settlement reflects reasonable responses to risk. AI too could provide some assessment of the confidence the judge should have in its predictions, information also suggestive of the range of plausible views about an appropriate settlement.

To be clear, for AI to play this sort of role would likely require procedural innovation. A court, for example, might require the parties to identify the most relevant legal precedents for evaluation by AI. If litigation has been pending for a while, that task might be easy or even unnecessary. The parties may have already cited the key case law. Or AI might be able to do legal research on its own.

Matters might become more complicated when it comes to the facts. The court might need the parties to present the evidence in a form that AI can evaluate—likely documents, including electronically stored information, as well as transcripts of testimony, such as from depositions. Again, if the litigation has progressed that task might be unnecessary. Summary judgment briefing might suffice. But it might not. The parties at summary judgment may not have addressed key issues for trial, or settlement may have come before briefing on summary judgment.

A new procedure might be necessary by which parties feed evidence to AI so it can do its job. It is beyond the scope of this Article to analyze whether Rule 23 empowers judges to use AI in this way (although there is precedent for courts scrutinizing class action settlements, including

[46] Fed. R. Civ. P. 23(e)(2).
[47] But see David Luban, Lawyers and Justice: An Ethical Study () (risk of double distortion rather than positions correcting each other).
[48] Citation on (professional) objectors.

by requesting additional information from the parties).[49] The main point, however, is that AI might be capable of helping judges exercise independent judgment in assessing a class action settlement.

E. Expected Value Mediation or Arbitration

AI could go beyond advising, advocating, and assisting. It could offer an alternative form of dispute resolution. We might call it expected value (EV) mediation or arbitration.

EV alternative dispute resolution would be most straightforward for monetary recoveries. It would involve AI calculating the expected value of the outcome of trial. To take a simplistic example, AI might determine that the plaintiff has a 50% chance of losing and a 50% chance of recovering $100,000. The expected value would then be 0.5 x $0 + 0.5 x $100,000 = $50,000. The parties might use that number as a guide for settlement in mediation. Alternatively, they might empower an arbitrator to impose the expected value of litigation to resolve a dispute, which we might call "Expected Value Arbitration" or "EVA."[50]

AI EVA might have numerous potential advantages over trial.[51] Those could include allowing parties to seek an independent judgment without the winner-take-all risks of resolution by a finder of fact,[52] minimizing harms from errors in legal decision-making,[53] and encouraging desirable expenditures on attorney's fees and costs, often at lower amounts than would traditional litigation.[54] Those benefits might be particularly great in class actions where there is a great deal at stake, the parties are likely to be averse to risk, errors may prove particularly costly, and litigation expenditures can be extraordinary.[55]

That said, AI EVA would give rise to some thorny issues. For example, should it consider the relative resources or the quality of counsel of the parties? Presumably, more expensive attorneys tend to skew the results of litigation in favor of their clients as compared to less expensive attorneys. Otherwise, we would have to assume that parties act systematically irrationally in paying for legal services. But we may be troubled if AI EVA were to adjust its analysis in light of the attorney's fees that the parties would anticipate expending in litigation. That could reward those with wealth in a way that is difficult to justify.[56] It could add yet one more advantage to the many that the "haves" hold over the "have nots."[57]

No doubt structuring AI EVA would entail other difficulties. But it nonetheless might resolve disputes with an efficiency and fairness that many litigants would find attractive.

---

[49] See, e.g., Judge Koh's denial of preliminary approval in High-Tech Cold Call?
[50] See Joshua P. Davis, Expected Value Arbitration, 57 Okla. L. Rev. 47 (2004).
[51] Id. at 70-106.
[52] Id. at 71-85.
[53] Id. at 85-94.
[54] Id. at 94-106.
[55] Citations.
[56] Davis, EVA at 119-121.
[57] Galanter.

V. Limits and Dangers of Robojudges

The above discussion identifies some legal tasks that we should expect AI to perform well. Its success is particularly likely if it improves greatly at abductive reasoning, even if it does not acquire purposive reasoning. Part V turns to a task that could be beyond the capacity of non-purposive AI: judging.

Part V.A explains why judges likely need to make value judgments to reach particular conclusions. Given our assumption that AI cannot make such judgments, AI would seem unable to fulfill the judicial role.

Part V.B then addresses a potential alternative endorsed by Eugene Volokh.[58] We might ask AI to use its power of prediction to write *persuasive* judicial opinions. Might such a robojudge perform as well as or even better than human judges? Part V.B offers reasons to doubt it would. It suggests that robojudges might be better than us at writing opinions that *seem* right but worse than us at writing opinions that *are* right. If so, robojudges might corrupt judicial decision-making rather than enhance it.

A. Judging and Pervasive Value Judgments

We have assumed that AI will not be able to make value judgments. The next issue is whether judges make value judgments when ruling in cases. That issue may seem to depend in part on an enduring controversy—the role of moral judgments in saying what the law is. That has been the primary debate in jurisprudence for over half a century.[59]

Fortunately, we need not resolve that debate to conclude that value judgments likely play a pervasive role in judging. Most jurisprudents acknowledge that moral judgments play a significant role in *creating* and *applying* the law, even if they do not or should not play a regular role in saying what the law is—in *interpreting* the law. Further, various legal values—including planning, authority, consistency, and predictability—may require judgments about ends, whether or not those values are moral.

Making Law. Consider the *creation* of new law. Some legal positivists deny that judges should make moral judgments in interpreting existing law. But they accept that judges, legislators, and the like should make moral or other value judgments when establishing *new* law. On that point, there is a widespread consensus.

That concession may seem relatively narrow. It is not. A reason is that there is no sharp distinction between making law and interpreting it. That is because uncertainty in law is a matter of degree. Generally speaking, there is no sharp distinction between, on one hand, extending existing law to fill gaps, resolve inconsistencies, and clarify ambiguities and, on the other, performing those same functions by creating new law. Put differently, judges will disagree about when they have merely interpreted existing law and when they have created new law. But little

[58] Eugene Volokh, Chief Justice Robots, 68 Duke L.J. 1135 (2019).
[59] See, e.g., Fuller, Hart, Dworkin, Coleman, Raz, Shapiro.

usually turns on that difference. Either way, judges tend to apply the law retroactively and speak as if they are simply discovering what that law is.[60]

Applying Law. A widespread consensus also exists that judges make moral or other value judgments in *applying* the law. A value judgment—even a moral judgment—may be embedded in a legal rule or standard. Contracts are unenforceable if they are unconscionable. Defendants are liable in tort if they do not take reasonable care. Plaintiffs may recover punitive damages if they prove malice. Assessing unconscionability, reasonable care, and malice involves value judgments, likely moral ones.

To be sure, jurisprudents disagree about whether those moral judgments are part of the law. So-called exclusive (or hard) legal positivists take the position that moral judgments are never part of the law.[61] They might say that is true even if the law relies on them—just as mathematics is not part of the law although the law at times uses it. In contrast, inclusive (or soft) legal positivists hold the view that the law can contain moral judgments (but that whether it does is ultimately a matter of pure social fact).[62] One might reasonably suspect that the disagreement here is more semantic than substantive. Regardless, legal positivists tend to accept that *applying* the law need not be a moral-free—much less a value-free—endeavor.

Again, this point may seem narrow. It too is not. No clear distinction exists between, on one hand, creating or interpreting law and, on the other, applying it. Consider so-called mixed questions of fact and law. Courts sometimes say that a mixed question exists when historical or primary facts are established or undisputed, but ultimate inferences and legal consequences are contested. [63] The line between a historical or primary fact and an ultimate inference, however, is fuzzy. Similarly, applications can shape rules and vice-versa. Factual scenarios can accrete into rules and rules can dissolve into factual issues, such that value judgments relevant to one can inform the other.

Interpreting Law. Jurisprudents also tend to agree that value judgments can inform legal interpretation. H.L.A. Hart acknowledged that purposive reasoning can play an important role in saying what the law is, although he denied that the purposes of the law are necessarily moral. He famously suggested, for example, that a key purpose of Nazi law was *evil*.[64] Joseph Raz argues that a distinctive understanding of legal *authority* forecloses moral judgments in legal

---

[60] Note unusual exception, including qualified immunity for a practical difference and talk of legal issues of first impression for a rhetorical difference.
[61] See, e.g., Raz and Shapiro.
[62] See, e.g., Hart and Coleman.
[63] See, e.g., *Pullman-Standard v. Swint*, 456 U.S. 273, 289 n.19 (1982) (mixed question of law and fact arises when the historical facts are established, the rule of law is undisputed, and the issue is whether the facts satisfy the legal rule); *Khan v. Holder*, 584 F.3d 773, 780 (9th Cir. 2009); *Suzy's Zoo v. Commissioner*, 273 F.3d 875, 878 (9th Cir. 2001) (mixed question "exists when primary facts are undisputed and ultimate inferences and legal consequences are in dispute"). Mixed questions of law and fact often require judgments about the values that animate legal principles. *See Smith v. Commissioner*, 300 F.3d 1023, 1028 (9th Cir. 2002).
[64] Reply to Lon Fuller in HLR debate.

interpretation.[65] Scott Shapiro contends that law is a *plan* (or plan-like norm) that provisionally resolves moral judgments and eliminates the need to revisit them in saying what the law is.[66] Evilness, authority, and planning are contestable and subtle values, even if not moral ones. Judges interpreting the law to serve those values would be expected to make judgments about them in legal interpretation.

To be sure, some readers may be skeptical about the legal positivism of Hart, Raz, and Shapiro. The law may not *always* serve moral purposes, but ideally it would *often* do so, at least in part. Further, whether a judge should consider morality in saying what the law is in any given setting would seem to depend in part on her judgments about political morality, including about the appropriate role for, say, an unelected judge in a representative democracy.

We should also note that authority and planning also seem like abstract moral values, as do internal consistency and predictability. Lon Fuller thus characterized such values as forming the internal morality of law.[67] Perhaps those values are not always moral. However, for legal interpreters attempting to fulfill their moral obligations—likely including many judges in representative democracies—they naturally would be interpreted as moral. If a judge, for example, attempts to abide by the moral responsibilities of her judicial office—assuming she has some—she would want to consider the moral force and nature of authority and planning. She thus would exercise moral judgment in deciding how she should implement those values and possibly how she should balance them against others, including achieving justice in particular cases.[68] From a moral perspective, it is hard to see how those issues could be anything other than moral ones.

In any case, moral or otherwise, authority and planning are values. The point is that although, for example, Justice Scalia endorsed different judicial value judgments than did Justice Cardozo—focusing more on consistency and predictability[69] than on societal changes and morality[70]— Scalia endorsed judicial value judgments nonetheless. As a result, AI cannot choose among them or fill in their content if it cannot engage in purposive reasoning.

B. Manipulation: What Seems Right, Not What Is Right

There is a strong case, then, that judging often involves value judgments, including likely moral ones. The next issue is whether robojudges might nonetheless be more effective than human judges at deciding cases.

---

[65] Raz.
[66] Shapiro, Legality (2011).
[67] See, e.g., Lon Fuller, The Morality of Law 46 (rev'd ed. 1969).
[68] Note Scalia's comment—which he later rescinded—that he would prove a faint-hearted originalist if that approach would allow whipping prisoners. Further note that he was willing to base his interpretive analysis on contested judgments about predictability and consistency, important and pervasive values in the law, and possibly moral values as well.
[69] E.g., A Matter of Interpretation; The Rule of Law as a Law of Rules.
[70] E.g., The Nature of the Judicial Process, 94-97, 133-34.

We do not have space to address that issue systematically. We can, however, consider an argument that Eugene Volokh made recently in "Chief Justice Robots."[71] Volokh contends that if AI is able to write more persuasive opinions than we can, we should accept AI as a judge. His position is characteristically thoughtful and forceful.

In responding to Volokh, we will develop a distinction that has broad application: between what *is* right and what *seems* right. That opens up the possibility that robojudges may be inferior to human judges, even if robojudges write more persuasive opinions than we do. Human beings may be better at determining what *is* right while AI may become better at predicting what will *seem* right. AI's opinions thus may be worse and yet more persuasive than ours.

Our analysis of these points will rely on some plausible assumptions that it will not defend. The first is that moral and other value judgments can be better and worse, maybe even right and wrong. The second is that moral and other value judgments matter—that we should act on better value judgments rather than worse ones. The third is that we have some capacity, however imperfect, to make accurate value judgments. The fourth is that we are also capable of erring in making value judgments, including if we are misled by self-interest or other biases.

To be sure, all of these assumptions are controversial. Credible philosophers would contest each one. But without them we would seem to have little prospect of determining what we morally should do or explaining why we should try to make sound value judgments at all. And, in any case, we lack space to justify these assumptions.

Why We May Write Better Judicial Opinions than AI Does

Let's begin with why our judicial opinions may be better than AI's. We have assumed that we can make value judgments but AI cannot. We have also concluded that value judgments are likely pervasive in judicial decision-making. AI, then, has to rely on our value judgments in writing opinions. That places AI at a disadvantage. Perhaps, however, AI can describe and predict our value judgments, and thereby make derivative value judgments of a quality similar to or even better than ours.

Changing Circumstances. For several reasons AI is unlikely to succeed in that task. First, circumstances change. AI often cannot apply old value judgments in a mechanical way to new settings. How old values apply to novel facts will not always be self-evident. The values may have to be clarified or refined. We can do that. AI cannot. It lacks the ability to form ends. As a result, AI will need new data from us to discern our views as the environment changes, whether those changes are, say, cultural, political, or, yes, possibly technological.

Changing Values. Second, values change. Of course, they may not always change for the better. But we lack a viable alternative to relying on evolving values. Otherwise, we might be stuck accepting that slavery, monarchy, and the like are as good as modern practices. We are the primary source of changes in values. AI can detect them only derivatively. That provides a second reason its value judgments will grow stale.

[71] Eugene Volokh, Chief Justice Robots, 68 Duke L.J. 1135 (2019).

Noise. Third, data about human value judgments are noisy. Past judicial opinions and other sources of law are tainted by biases, base instincts, psychological desires, and related products of motivated cognition. We have some hope of distilling our value judgments from such noise. With diligence, self-discipline, and candor, we may be able to disentangle what we think is right from what we want to believe.

True, our efforts along these lines are likely to be flawed. But it is not clear how AI can distinguish value judgments from biases at all. Beliefs do not come pre-labeled. Only substantive value judgments enable us to distinguish one from the other. As we have noted, AI cannot make substantive value judgments. It has to reach conclusions about values based on what we say and do. Its inferences will thus be tainted to the extent our statements and actions are. We, in contrast, may be able to discern the signal of our insights about values from the noise of our motivated cognition, however imperfectly we do so.

Why AI Nevertheless May Write More Persuasive Judicial Opinions than We Do

Our advantage at making sound value judgments could lead us to write more persuasive judicial opinions than AI does. But it may not. The silver tongue of the devil may convince us more effectively than honesty of our better angels.

Put less poetically, AI may win us over by telling us what we want to hear rather than what we should hear. Some of our value judgments are uncomfortable. They can reveal that we have been acting inconsistently with what we consider best on reflection. Perhaps we have interpreted the law or applied it in ways that reflect systemic biases. Perhaps we have adopted flattering views about ourselves that cannot survive scrutiny. We may resist those and similar possibilities and, as a result, prefer judicial opinions that deftly rationalize what we have done wrong, or what we would like to be true, to judicial opinions that in a deeper sense are right.

Moreover, it is often difficult to make clear statements about the law or its application when we are recalibrating. We do not always see how a new approach will play out. That is one of the reasons courts at times emphasize that they resolve disputes one at a time and that statements about circumstances not before a court are dicta. It can take a while for legal change to cohere. Until it does, the opinion that is most persuasive—perhaps because it offers an orderly statement of the law—may not be the best one.

Further, if AI is directed to write judicial opinions that judges or others will find most persuasive, we should expect it to *exploit* weaknesses in how we think. It will not do so out of some improper motivation. It has no motivations. To function, it will need data about what judges and others find persuasive. Those data would presumably embody all sorts of inclinations that judges and others may try to resist. A robojudge, however, will lack the capacity to distinguish sound arguments from manipulative or dangerous ones. It will by its nature discover and take advantage of ingenious ways to make illegitimate arguments seem legitimate. That could make AI judicial opinions more persuasive than ours even though—perhaps *because*—they are inferior.

Why AI May Corrupt Judicial Decision-Making

This last point suggests the possibility that robojudges could *corrupt* judicial decision-making. That could occur in various ways, each potentially compounding the others.

Staleness. Robojudges could deprive themselves of data. If they take over all or most of the judiciary, they will no longer have recent human judicial opinions from which to detect patterns. As our circumstances and values change, AI opinions are likely to grow progressively staler. They may no longer reflect modern society and its beliefs.[72] Of course, we could potentially detect such staleness and compensate for it, maybe by infusing the body of AI opinions with human ones. The other sources of corruption discussed below, however, may interfere with our capacity or motivation to do so.

Excessive Deference. We may confuse AI's acumen at persuasive legal analysis for sound value judgments. AI may become extraordinarily skilled at interpreting and applying the law in ways we find credible. That may blind us to the implicit value judgments it makes in assessing how to persuade us. One way to understand this phenomenon is as a variation on the halo effect: our tendency to ascribe to people positive qualities that they do not have because of positive qualities that they do have. We may believe robojudges make sound values judgments because of their skill at opinion writing. A result is that we may miss that the law is listing away from what is right.

Atrophy. Over time we might lose the *ability* to engage in effective legal reasoning. It is a skill. Without practice, skills deteriorate. Without a large stock of human judges—or at least human lawyers—we should not assume that we will remain capable of overseeing or evaluating robojudges.

Complacency. We may also lose the *motivation* to engage in judicial reasoning. It is hard work. It can be stressful. Conscientious judges often struggle in choosing between competing arguments or developing their own independent views. If robojudges *seem* to be doing the work of judging well—even if a careful analysis would reveal that they are making serious and accumulating errors—we may not undertake the arduous work necessary to discover the problem and correct course. Especially if we become weaker at judicial reasoning and less motivated to do it, we may find ourselves susceptible when AI appeals to our base instincts and habits.

Distortion of our Values. AI may even write judicial opinions that are *designed* to shape our preferences so that they are more predictable. That could taint any feedback loop we develop in an effort to ensure the ongoing quality of AI opinions. In this regard, recall how AI directed social media users to links that would shape the users' views so that their search habits became more predictable. Stuart Russell suggests that as an explanation for the way in which social media fosters extreme political views.

---

[72] For an argument about a different way in which AI may stunt development of the law see Daniel Maggen, Predict and Suspect: The Emergence of Artificial Legal Meaning 23 N. Car. J. of Law & Tech. __ (2021) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3904262&dgcid=ejournal_htmlemail_jurisprudence:legal:philosophy:ejournal_abstractlink).

A similar phenomenon could occur in law. AI might end up writing judicial opinions designed to alter our views and preferences so that it can more effectively persuade us. Rather than AI opinions merely reflecting what we find persuasive, they also might *shape* what we find persuasive.

To be sure, the above analysis does not prove that robojudges would be more persuasive opinion writers than human judges or that robojudges would writing worse opinions than human judges. There are other possibilities. One is that, all things considered, human judges remain more persuasive than robojudges. Another is that AI, on the whole, will write better opinions—perhaps because its superior technical abilities will more than compensate for its limitations regarding value judgments. Still, we have reason for caution despite AI's potential capacity to persuade—indeed, potentially *because of* its capacity to persuade. AI may turn out to be a devil with a silver tongue.[73]

We can encapsulate some of the above analysis by considering two passages in Volokh's article. The first quotes Justice Kagan when she "described the shift from Solicitor General to Supreme Court Justice as shifting 'from persuading nine [Justices] to persuading eight.'"[74] Volokh uses the quotation as evidence that judges write opinions to persuade each other, which seems reasonable.

Note, however, that judges first must decide how they think a case *should* be decided. Justice Kagan presumably would never write an opinion *only* to maximize its odds of winning the other Justices' votes. That would be absurd. If all Justices did that, they would end up in an infinite regress, much like two mirrors facing each other. Each would write an opinion reflecting the anticipated views of the other Justices, which would reflect the anticipated views of the other Justices, ad infinitum. None of them would be making any direct judgments about the law or the facts or how the two relate.

Another way to illustrate this point is by responding to one of Volokh's rhetorical questions: "What more can we reasonably ask of an opinion drafter—human or AI—than the

---

[73] Volokh suggests the possibility that we could program robojudges to make persuasive arguments about what is wise or compassionate—and not just about what is legal—if that is what we want. Cite. Note, however, that the same points made in the text apply to wisdom and compassion as to law. AI may be more effective at determining what seems wise or compassionate than what is wise or compassionate. A common thread is that wisdom and compassion—like law—are what philosophers sometimes call thick ethical concepts (or perhaps we should say thick normative concepts, to capture other values). *See, e.g.*, PUTNAM, THE COLLAPSE OF THE FACT/VALUE DICHOTOMY AND OTHER ESSAYS 34-43. The point is: we can no more say with determinacy what is wise or compassionate without making value judgments than we can say what is legal. (One might similarly question any sharp distinction between what is wise or compassionate and what is legal.)

[74] Volokh at 1149-50 & n. 49 (quoting Phil Brown, Associate Justice Elena Kagan Visits NYU Law, NYU L. Commentator (Apr. 5, 2016), https://nyulawcommentator.org/2016/04/05/associate-justiceelena-kagan-visits-nyu-law [https://perma.cc/3N32-8KAR] (quoting Justice Kagan).

production of opinions that a blue-ribbon panel of trained observers will accept over the alternatives?"[75] But we can—and do—ask more of some human opinion drafters. We ask judges to try to get their decisions *right*. If AI cannot do that, it may be a poor substitute for us.

VI. Conclusion

AI has made tremendous strides at deductive and inductive reasoning. It may in the not-too-distant future improve similarly at abductive reasoning—which could include the kind of common sense that figures prominently in lawyering and judging. If so, that might greatly expand the role of AI in litigation in general and in complex litigation in particular. It could advise us, advocate for us, help judges assess class action settlements, and propose or impose compromises to resolve legal disputes.

But that does not mean that AI would be able to serve as an effective judge. There is good reason to believe doing so requires a capacity to make reliable judgments about morality or other values. There is also good reason to believe that AI will not be capable of making those reliable judgments.[76] That may be true even if we find opinions drafted by robojudges more persuasive than ones drafted by human judges. We should take care about ceding the judiciary in whole or in part to AI. Doing so might corrupt our legal system—rendering the law progressively less just over time.

[75] Volokh, at 1154.

[76] The analysis has assumed that AI will not acquire consciousness. It may. I will address reasons to doubt that conscious AI would be capable of reliable judgments about morality or values in JOSHUA P. DAVIS, UNNATURAL LAW: AI, CONSCIOUSNESS, ETHICS, AND LEGAL THEORY (forthcoming Cambridge University Press 2022/23).