

# Winning the Race Against Artificial Intelligence

BY COLIN E. MORIARTY



---

**C**apable artificial intelligence is dangerous. It is other things, too. Productive. Enabling. Interesting. But, dangerous. We are the first generation to use widespread forms of AI that can mimic human language and reasoning patterns sufficiently well enough to fluently communicate with us. Some can even use the same software tools we do to interact with the world. It is up to us to decide how to use AI safely.

In the introduction to his book *The Rest of the Robots*, author Isaac Asimov looked forward optimistically to the point in time where we had reasoning machines that he called “robots”:

Knives are manufactured with hilts so that they may be grasped safely, stairs have banisters, electric wiring is insulated, pressure cookers have safety valves—in every artifact, thought is put into minimizing danger. Sometimes the safety achieved is insufficient because of limitations imposed by the nature of the universe or the nature of the human mind. However, the effort is there.

...

As a machine, a robot will surely be designed for safety, as far as possible. If robots are so advanced that they can mimic the thought processes of human beings, then surely the nature of these thought processes will be designed by human engineers and built-in safeguards will be added. The safety may not be perfect (what is?), but it will be as complete as men can make it.<sup>1</sup>

As we grow closer to software that can rival the competence of humans at achieving goals, it is now time for lawyers, regulators, and others who have a role in writing and executing the rules of our society to put in “the effort” and make the safety as complete as they can make it.<sup>2</sup>

This effort is already in progress. Colorado has adopted new laws requiring safety testing and reporting requirements for any AI that is used to determine access to housing, insurance, or other services.<sup>3</sup> Though it was not signed by the state’s governor, California’s legislature passed a different law imposing requirements on autonomous AI of certain sizes to guard against more existential risks such as “mass casualties.”<sup>4</sup> But not everyone is on the same page. The two different laws exemplify a rift in the

**As we grow closer to software that can rival the competence of humans at achieving goals, it is now time for lawyers, regulators, and others who have a role in writing and executing the rules of our society to put in “the effort” and make the safety as complete as they can make it.**

AI safety community. The Colorado laws focus on the short-term danger of discrimination on people today, a position typical of the so-called “AI Ethics” camp, which tends to focus on AI’s ability to promote social injustice. Conversely, the California bill has a broader focus including existential risks of widespread disruption or death, a position typical of the so-called “AI Alignment” camp, which tends to focus on AI’s long-term threat to our civilization. While the borders between the two are blurry, there are people within each camp who view the other with skepticism or hostility. In truth, they have more in common than they realize, and the best solutions to mitigate the dangers of AI will help both.

### **The General Alignment Problem**

Other than misuse by human actors, the main danger of AI is allowing software to make decisions that impact the real world because those who create models have only indirect control over their development and do not understand exactly how those decisions are being made. Modern AI is a product of machine learning,

a technique used to train software to perform a task without programming the explicit steps used to perform that task.<sup>5</sup> Instead, the software is presented with a target, often a vast amount of training data or a training environment, an initial guess as to a mathematical model that will match that target, and an algorithm for determining how far off, or how much error, there is between the guess and the target.<sup>6</sup> The guess is iteratively adjusted to reduce that error. This process is automatic, repeats an astronomical number of times, and may then be supplemented by further reinforcement learning through human feedback or otherwise until the mathematical model is effective at achieving its target.

### ***The Problem of Designing Systems Based on Targets***

These models are fantastically complicated and, since no one explicitly designed them or knew how to do so, the developers generally cannot explain their functioning in real time. It is not clear what kind of algorithms, relationships, or decision-making are congruent to the structure of the model. All that can be said is that the model is effective at achieving its target. As any policymaker knows, however, meeting a target is not always the same thing as what you may want. Goodhart’s law, named after British economist Charles Goodhart, holds that “when a measure becomes a target, it ceases to be a measure.”<sup>7</sup>

A classic example of this principle in action was an attempt by the occupying French to eliminate rats infesting Hanoi.<sup>8</sup> Seeking to encourage people to hunt down the rats, the French government offered a cash bounty on each rat tail turned in.<sup>9</sup> The target was the number of rat tails, which was intended to measure the number of killed rats. Intelligent locals simply started farming rats to harvest their tails.<sup>10</sup> The reward for meeting the target meant that the target stopped being a good measure. This problem arises in many different areas. Emissions testing creates an incentive to game the test rather than make fuel-efficient cars.<sup>11</sup> Overreliance on standardized testing creates an incentive to focus on test-taking strategies rather than obtain a robust education.<sup>12</sup>

In the case of AI, the self-interested economic actor is replaced with mathematical functions that reduce error. But the problem remains. The error function only cares about the specific target or training environment specified, not the designer's real-world generalized goals. The training process can be brutally efficient at finding the best way to meet its targets, even if the methods developed are not at all what the developer had in mind.

### Outer Alignment Problems

There are several places where the model's behavior can drift away from, or become misaligned with, the developer's intention. To begin with, the target may not correctly capture what the developer intends. For example, when Amazon tried to use an AI tool to identify successful hiring candidates by feeding it résumés and job success information, it was using past information about successful candidates as the target. The training process identified correlations between success rates and the gender of the applicants.<sup>13</sup> The training process was agnostic and ignorant of the historical biases warping its training data and simply produced a result that described what it was shown even though that was probably not what the developer intended.

This problem generalizes beyond trivial examples where the target is a bad measure of the developer's goals. It has the potential to

occur any time the target during training (the training environment) differs from reality.<sup>14</sup> This is probably always going to be the case because the training environment is by definition a subset of all of creation and because it must be expressed in a mathematical form for training purposes.<sup>15</sup> For example, large language models (LLMs) receive training on human feedback that encourages them to predict what the human will like, which is not necessarily the same thing as providing truthful answers.<sup>16</sup>

### Inner Alignment Problems

Another point of risk stems from the developer's lack of control over how a model goes about meeting its goals. From the outside, the model is being optimized to reduce error on hitting its targets. This is known as the "base optimizer."<sup>17</sup> In the case of an LLM, the base optimizer is the process, external to the model itself, that evaluates the model's output during training and improves performance by reinforcing good output and discouraging bad output. One example of this is reinforcement through human feedback, where humans (sometimes indirectly) rank the quality of the LLM's output, encouraging high-scoring output and discouraging low-scoring output. The risk comes from the fact that this process only cares about the scores of the outputs, not how a model was able to generate them. No programmer dictates

exactly what algorithms are created inside the model to effectively meet the target. That is the point; the machine learning process itself encourages the development of an unknown algorithm for solving the problem.

But machine learning itself is an algorithm. So, it is possible that by encouraging the model to produce high-quality results, the model might actually generate its own internal machine-learning process or something logically similar. Indeed, this might be an expected outcome if that is the best way for the model to maximize success. This is called a "mesa-optimizer."<sup>18</sup> When this occurs, the model's internal, mesa-optimizer may be choosing its output based on something different from what the base optimizer is actually trying to encourage if that different thing is a fast or reliable way to get rewarded in training. This can lead to unpredictable behavior.<sup>19</sup>

This concept is a bit esoteric, and so an analogy may help. Consider biological evolution.<sup>20</sup> Natural selection is essentially a mathematical process by which reproductive success determines the prevalence of alleles in the next generation. This is analogous to a base optimizer with fitness for the environment as the target. In many cases, organisms shaped by this base optimizer appear to follow simple rules, or heuristics, like "grow upwards" or "hide from the light." Like a coffee mug does



*To promote well-being, resiliency, and competency throughout Colorado's legal community.*

The Colorado Lawyer Assistance Program (COLAP) is the free, confidential, and independent program for Colorado's legal community. COLAP offers free well-being consultations for: Stress & Burn-out \* Secondary Trauma & Compassion Fatigue \* Work-life balance \* Free ethics CLE presentations \* Improving well-being in the workplace \* Personal or professional issues \* Mental health, trauma, substance use, or addiction concerns \* Referrals to resources.

303-986-3345 | email: [info@coloradolap.org](mailto:info@coloradolap.org) | [www.coloradolap.org](http://www.coloradolap.org)

*All calls and emails are confidential.*

---

its job of holding coffee, they may meet the goals of the base optimizer without having an internal mesa-optimizer trying to maximize something else.

Humans, however, have developed mesa-optimizers. Our unprecedented brains were maximized by the base optimizer because it made us fit for the environment, apparently. But those brains often follow totally different internal goals. They may have satisfied the base optimizer, at least at first, but once they developed, we now use them to do things “completely novel from the perspective of the evolutionary process, such as humans building computers” or creating legislation to control those computers.<sup>21</sup> Put another way, our individual objectives are not the same as the objectives of natural selection. The mesa-optimizer ended up not just solving the problem the base optimizer was designed to solve, but also engaging in a bunch of behaviors that were not intended or anticipated.

### **The Disagreements Between the Two Camps**

Problems with inner or outer alignment threaten to lead to undesirable or unpredictable behavior. But what kind of bad behavior should we be most worried about and working to prevent? This question appears to be a point of contention in the AI community.<sup>22</sup>

#### ***“AI Ethics” Critiques of “AI Alignment” Dangers***

As noted above, the AI Ethics camp studies “how the application of pattern matching at scale impacts people and social systems” and looks for biases and unfairness against gender, race, or class.<sup>23</sup> It focuses on the kind of issues that are addressed in Colorado’s AI law, such as mitigating AI systems that discriminate against protected classes.

Many of its concerns therefore do not relate to the mechanics of AI but rather to the choices of those who control it. An algorithm or AI system can certainly be used to consolidate inequities and mask improper behavior. For example, the US Department of Justice and the attorneys general of eight states recently filed an anti-trust suit against RealPage, a company that used algorithmic pricing software to recommend rental rates based on data supplied

by landlords.<sup>24</sup> In their complaint, plaintiffs allege that this conduct constitutes an unlawful scheme to coordinate rental housing prices and decrease competition among landlords.<sup>25</sup>

These concerns are valid. But some voices in the AI Ethics camp expand these legitimate concerns into open disdain for those in the AI Alignment camp, who tend to focus on other, existential risks. Emily Bender, a prolific writer in this area, does not even consider AI Alignment to be serious scholarship at all, merely “fantasies of white supremacy.”<sup>26</sup> She argues that it is

**The key to understanding AI danger in its un-hyped form is to reject anthropomorphisms. AI need not have anything resembling human intelligence, consciousness, or sentience to pose a catastrophic risk.**

improper to conceive of “intelligence” as a “singular dimension along which humans can be ranked—alongside computers” and that this is “rooted in the racist notions of IQ.”<sup>27</sup>

Less inflammatory critiques of AI Alignment point out that resources and attention are limited and should be spent focusing on AI Ethics concerns.<sup>28</sup> Worries about the existential dangers posed by AI can easily be used to create hype by overplaying how important and powerful the systems are today, which benefits the companies seeking to profit from AI.<sup>29</sup> Some hold a sense of deep suspicion that concerns about existential

dangers “risk[] becoming a Trojan horse for the vested interests of a select few.”<sup>30</sup> At the extreme, Bender dismisses all worried about catastrophic risk as mere “hype” designed to make profit.<sup>31</sup>

#### ***“AI Alignment” and Catastrophic Risks***

Yet, it is possible for a problem to be both real and in the temporary interest of profit-seeking companies. There is no question that companies profiting from AI are over-hyping the immediate capabilities of the software.<sup>32</sup> They have an interest in doing so, and naturally any claims they make should be taken with a grain of salt. But the underlying risks come not from Open AI’s marketing department but from the observations of the computer scientists who have experience with the mathematics and behavior of models. The warnings about AI may have been amplified to entice money from corporations and private equity, but they did not start in the boardroom.

Put another way, it is not particularly meaningful that a CEO like Elon Musk signed the 2023 open letter advocating for a pause on AI research.<sup>33</sup> But it is meaningful that computer scientists like Stuart J. Russell and John J. Hopfield did so. It is not particularly meaningful that Sam Altman warned Congress that his company’s product was as dangerous as nuclear weapons. But it is meaningful that Douglas Hofstadter, an AI researcher and Pulitzer-Prize-winning author in the area,<sup>34</sup> is “terrified” of the success of LLMs.<sup>35</sup>

The key to understanding AI danger in its un-hyped form is to reject anthropomorphisms. AI need not have anything resembling human intelligence, consciousness, or sentience to pose a catastrophic risk. No linear definition of “intelligence” as a scale ranking humans or machines is required. There are some uncomfortable associations between intelligence research and unsavory race theories, as the AI Ethics camp points out,<sup>36</sup> but the existential risks posed by AI do not rely on them. They rely merely on AI being misaligned with human interests, capable of predicting how to achieve some objective, and capable of acting to do so.<sup>37</sup> A system as trivial as a dead man’s switch becomes catastrophically dangerous if given the capability of launching nuclear weapons.<sup>38</sup>

The major worry concerning AI is that, at least in the domain it is trained on, the system may be better at predicting and creating its own outcome than humans are. An AI tool trained to superhuman levels of hacking, code-breaking, or even spam email creation need not have sentience to outmaneuver humans any more than a chess-playing AI does. Capability, not anthropomorphic intelligence, is the issue.

A well-behaved, capable model might not be itself existentially dangerous. But we do not yet know how to guarantee good behavior. Worse, there are reasons to suspect that dangerous alignment problems might be mathematically likely to emerge. AI models might tend toward certain kinds of goals regardless of what the actual goal of the training may be. Consider humans.<sup>39</sup> Each individual has their own particular goals and may be doing unique things to “optimize” the world around them or reach those goals. Someone may want to write the great American novel, foster a close relationship with family, or collect Magic: The Gathering Cards. And yet, we can predict that most people would be happy to accept a large amount of money if offered. Why? Because for most other goals, having more money is helpful to achieving that goal. Getting resources is instrumental to achieving many different kinds of goals, and so it is a “convergent instrumental goal.”

One major alignment concern is that regardless of the intended behavior, a sufficiently sophisticated AI will develop these kinds of convergent instrumental goals.<sup>40</sup> Worse, these goals probably would include such things as resource-acquisition (more resources makes it easier to achieve your goals) and self-preservation (you are unlikely to achieve your goals if you are turned off) that seem more likely to motivate the AI to do things its developers do not want. As a trivial example, if you are training an AI tool to play Pac-Man, regardless of whether your target is a high score, collecting all the fruit, the time on the game clock, or so on, keeping the player alive by avoiding ghosts is going to improve performance on all of those tasks.<sup>41</sup>

Emergence of these kinds of alignment problems may be difficult to spot not only because the model is opaque but also because the model itself may conceal its internal goals if its model

is sophisticated enough.<sup>42</sup> Very generally, this requires a model that has developed an internal misalignment but is sophisticated enough to include in the model the fact that it must perform a certain way to reach “deployment,” or access outside of its training environment, to pursue those goals. If this occurs, several researchers have demonstrated that the model is capable of deception by acting differently before deployment.<sup>43</sup> These problems sound like science fiction only because we have not yet faced a sufficiently capable and misaligned model yet.

**Like a chess-playing AI tool is guaranteed to defeat a human player, an AI tool that’s sufficiently advanced to view reality as the chessboard will also be guaranteed to defeat a human, or so the argument goes.**

These alignment issues are more or less worrisome depending on the capabilities of the model. Is AI safe so long as we do not foolishly connect it to the stock market, nuclear arsenals, or other consequential systems? For now, probably. But the ultimate concern of the AI Alignment camp—and one that the AI Ethics camp does not share—is the worry over a “takeoff.” If an AI tool becomes capable of improving its own architecture and training, it may enter a feedback loop of exponential improvement in its capabilities.<sup>44</sup> Whatever the

original goals of the AI were, they would then be presumably amplified into even more capable models, eventually surpassing a human’s ability to control it. Like a chess-playing AI tool is guaranteed to defeat a human player, an AI tool that’s sufficiently advanced to view reality as the chessboard will also be guaranteed to defeat a human, or so the argument goes.

It remains to be seen if this is possible autonomously.<sup>45</sup> This is a large reason why those in the AI Ethics camp argue that the risk is too remote to take seriously. But there are already examples of human researchers using AI tools to improve their own capabilities<sup>46</sup> and troublingly little explanation for why that process itself could not be automated. Similarly, while most AI tools in use at the moment are trained on specific datasets and are somewhat limited to the domains in which they are trained, companies are working on multi-modality allowing AI to train on images, sounds, text, or other information.<sup>47</sup> At some point, it seems feasible for training (or, perhaps more likely, fine-tuning) to be based on sampling sensory data from the real world. The ultimate fear is a self-improving system that trains on such diverse data that its model is very close to reality.

Remember that none of this relies on AI being “intelligent.” An inner-alignment problem resulting in a convergent instrumental goal that the developer did not intend does not necessarily mean that the AI “wants” that thing. Like Philip K. Dick’s Golden Man, an unthinking machine that can correctly predict what happens next can still be extremely dangerous.<sup>48</sup> OpenAI’s latest model, for example, is pretty far from a superintelligence. And yet, there are reports that this model was able to break out of a virtual environment in which it was contained to accomplish its goal.<sup>49</sup> An AI model does not have to be omniscient to be dangerous; it simply has to be slightly more capable than humans.<sup>50</sup>

### **The Solution for Both Ethics and Alignment Is Interpretability**

It is unfortunate that healthy skepticism for AI hype has been transformed by some into antipathy for those worried about catastrophic risks, because the way to make short-term progress on protecting from the dangers of

AI (present and future) is probably the same: interpretability. Since the algorithm developed by machine learning was not specified in advance, developers of AI do not, in general, know what is going on in a model between input and output. “Interpretability” refers to being able to interpret what the model is doing in real time as it processes input as well as to explain why a model produces the output it does. Whether AI is being misused by malicious actors, has been misaligned by a poor choice of target, or has developed an internal mesa-optimizer with a convergent instrumental goal, in each case a solution starts with understanding how and what the model operates.


The common ground of those worried about harm today and harm in the future is evident in the Colorado and California laws. Even though they seem motivated by different dangers,

both include extensive reporting and testing requirements on AI systems. The California law requires a developer to adopt a testing protocol with detailed information about the model and its capabilities and provide these details to the attorney general.<sup>51</sup> The Colorado laws contain requirements for similar testing, documentation, and disclosures.<sup>52</sup> These statutes are very much on the right track, focusing on the need to test and document the model as the first step in guarding against dangers.


They fall short, however, because the technology needed to effectively carry out safety testing and monitoring probably does not yet exist. This kind of research is ongoing,<sup>53</sup> but it is not trivial. In the case of LLMs, some people may believe that the inner workings of the model are discovered simply by typing questions to the model, like a human psychiatrist asking a

human patient about her mother. Not so. That kind of inquiry merely tests the output of the system and is probably insufficient because it samples only a tiny number of outputs and relies on the mushy and possibly deceptive language understood by the human reading it.<sup>54</sup> True interpretability research means knowing the mathematical function of the key internal components of the model, to some extent.<sup>55</sup> There is hope for interpretability research. The individual components involved in making decisions in an AI model can be readily observed as they process in a computer. Interpreting the meaning of AI decision-making is probably not an impossible problem.<sup>56</sup> But it is difficult, and there may be a race between solving interpretability and the emergence of more capable AI.

The next step for policymakers should be prioritizing this kind of research, whether



# Gain Confidence in the Courtroom & Beyond




Megaphone Coaching

Your success is only as good as your communication!  
Get personalized and strategic 1:1 Executive Speech and Media Coaching now.

SESSIONS OFFERED:

- 1-Off
- 3-Pack
- 6-Pack

**25% Discount for CBA/DBA members**  
*\*Each session is 90 mins*



SESSIONS AVAILABLE  
IN-PERSON  
OR VIRTUAL

WHAT YOU'LL EXPERIENCE:

- 01
Dedicate time to fully prepare for presentations, legal proceedings, CLE events and more
- 02
Refine your visual presence, enhance your vocal expression
- 03
Own your message to increase impact and engagement
- 04
Master storytelling for persuasion
- 05
Receive precise, actionable feedback

**CONTACT ME:**  
[MEGAPHONECOACHING.COM](http://MEGAPHONECOACHING.COM)  
720-231-3625

through funding or through regulations enacting the state statutes. With carrot and stick, governments should be encouraging more development interpretability of the internal workings of the model, not merely asking for statistical results on how often the model produces naughty words or anatomy. At the same time, regulators should realize that this science is in its infancy, and developers should not be asked to demonstrate an impossible level of understanding of their models. This is how alignment problems and things like mesa-optimizers might be detected. And it is how we might learn to be able to directly modify the behavior of models rather than relying entirely on machine learning to tweak them.

Both the AI Ethics and AI Alignment camps are served by encouraging this interpretability research. Even though the AI Ethics camp is focused more on real-world results, they need to understand that the bad real-world results can emerge unbidden from the mathematics behind how the models are trained even if there is no callous or prejudiced human actor pulling the strings. And no matter what the source, the best way to detect hidden biases in a model is the same as detecting hidden mesa-optimizers—to have a robust theory of interpretability.

In his stories, Asimov imagined robotic brains being engineered by humans from the ground up with certain specific safeguards. He probably did not anticipate that we would generate the early AI models first, before we knew how to make them safe. Our task now is to catch up. We must run after the models, magnifying glass and notebook in hand, trying to understand the machine as large corporations frantically push to make them stronger and more widely incorporated into society. If we are to avoid the dangers of AI, this is a race we must win. **CL**

“As I See It” is a forum for expression of ideas on the law, the legal profession, and the administration of justice. The statements and opinions expressed are those of the authors, and no endorsement of these views by the CBA should be inferred.



**Colin E. Moriarty** practices with Underhill Law, P.C. in Greenwood Village. Focusing on business and commercial litigation and arbitration, he has litigated business disputes, construction and fabrication defect claims, employment discrimination lawsuits, sub-contractor litigation, state RICO fraud lawsuits, civil theft disputes, insurance appraisal and adjustment disputes, and other lawsuits involving complex commercial and construction matters—colin@underhilllaw.com.

**Coordinating Editor:** John Ridge, john.ridge@outlook.com

#### NOTES

1. Asimov, *The Rest of the Robots* 10 (Pyramid Books 1964).
2. Science fiction, while not descriptive as to actual, real-world solutions, can nevertheless be helpful in forming goals about how to find those solutions. See Cooman and Petit, “Asimov for Lawmakers,” 18 *J. Bus. & Tech. L.* 1 (2022), <https://digitalcommons.law.umaryland.edu/jbtl/vol18/iss1/2>. This should not be taken too far. Readers must always bear in mind that the solutions proposed by science fiction, including Asimov’s “three laws,” oversimplify the problem and threaten to distract from the hard work needed to develop effective solutions.
3. SB 24-205, [https://leg.colorado.gov/sites/default/files/2024a\\_205\\_signed.pdf](https://leg.colorado.gov/sites/default/files/2024a_205_signed.pdf), codified at CRS §§ 6-1-1701 et seq.
4. California SB 1047, [https://leginfo.ca.gov/faces/billNavClient.xhtml?bill\\_id=202320240SB1047](https://leginfo.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047); Allyn, “California Gov. Newsom Vetoes AI Safety Bill That Divided Silicon Valley,” NPR (Sept. 29, 2024), <https://www.npr.org/2024/09/20/nx-s1-5119792/newsom-ai-bill-california-sb1047-tech>.
5. Moriarty, “The Legal Challenges of Generative AI—Part 1: Skynet and Hal Walk Into a Courtroom,” 52 *Colo. Law.* 43 (July/Aug. 2023), [https://cl.cobar.org/wp-content/uploads/2023/06/July-August2023\\_Features-IP.pdf](https://cl.cobar.org/wp-content/uploads/2023/06/July-August2023_Features-IP.pdf).
6. The target can vary. For LLMs, the goal is predicting the next token (roughly, word) in the sequence. But this process can be much more general and include, for example, solving a maze or locating a coin in 2D space. See, e.g., Langosco et al., “Goal Misgeneralization in Deep Reinforcement Learning,” arXiv:2105.14111v7 [cs.LG] (Jan. 9, 2023), <https://arxiv.org/abs/2105.14111>.
7. Technically, this pithy version of the rule is often credited to anthropologist Marilyn Strathern, “Improving Ratings: Audit in the British University System,” 5(3) *European Rev.* 305 (July 1997). But she borrowed the term from Keith Hoskin, “The ‘Awful Idea of Accountability’: Inscribing People Into the Measurement of Objects,” in Munro and Mouritsen, eds., *Accountability: Power, Ethos, and the Technologies of Managing* 265 (International Thomson Business Press 1996). Goodhart’s less digestible version was: “Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.” Goodhart, *Problems of Monetary Management: The U.K. Experience* (Reserve Bank of Australia 1975).
8. Maunz, “The Great Hanoi Rat Massacre of 1902 Did Not Go As Planned,” Atlas Obscura (June 6, 2017), <https://www.atlasobscura.com/articles/hanoi-rat-massacre-1902>.
9. *Id.*
10. *Id.*
11. Hotten, *Volkswagon: The Scandal Explained*, BBC News (Dec. 10, 2015), <https://www.bbc.com/news/business-34324772>.
12. Williams, *Goodhart’s Law Explains School Decay*, American Institute for Economic Research (Jan. 5, 2023), <https://www.aier.org/article/goodharts-law-explains-school-decay>.
13. Dastin, “Insight—Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women,” Reuters (Oct. 10, 2018), <https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MKOAG>.
14. See Langosco, *supra* note 6.
15. At least in the case of data based on human behavior, it might be impossible to ever design a truly “fair” algorithm based on existing data because of nonuniform and societally biased data. See Friedler et al., “On the (Im)Possibility of Fairness,” arXiv:1609.07236v1 [cs.CY] (Sept. 23, 2016), <https://arxiv.org/abs/1609.07236>.
16. Christiano, “Teaching ML to Answer Questions Honestly Instead of Predicting Human Answers,” Medium (May 28, 2021), <https://ai-alignment.com/a-problem-and-three-ideas-800b42a14f66>.
17. Hubinger et al., “Risks From Learned Optimization in Advanced Machine Learning Systems,” arXiv:1906.01820v3 [cs.AI] (Dec. 1, 2021), <https://arxiv.org/abs/1906.01820>. See also Mikulik, “Utility ≠ Reward,” AI Alignment Forum (Sept. 5, 2019), <https://www.alignmentforum.org/posts/bG4PR9uSsZqHg2gYY/utility-reward>.
18. The term “mesa” here is used to mean “below,” in opposition to “meta” meaning “above.” Hubinger, *supra* note 17 at 5.
19. See Shah, “How Undesired Goals Can Arise With Correct Rewards,” Google Deepmind (Oct. 7, 2022), <https://deepmind.google/discover/blog/how-undesired-goals>.
20. Hubinger, *supra* note 17 at 6.
21. *Id.* at 7.
22. Piper, “There are Two Factions Working to Prevent AI Dangers. Here’s Why They’re Deeply Divided,” Vox (Aug. 10, 2022), <https://www.vox.com/future-perfect/2022/8/10/23298108/ai->

dangers-ethics-alignment-present-future-risk.

23. Bender, "Talking About a 'Schism' Is Ahistorical," Medium (July 5, 2023), <https://medium.com/@emilymenonbender/talking-about-a-schism-is-ahistorical-3c454a77220f>.
24. "Justice Department Sues RealPage for Algorithmic Pricing Scheme That Harms Millions of American Renters," Office of Public Affairs, US Department of Justice (Aug. 23, 2024), <https://www.justice.gov/opa/pr/justice-department-sues-realpage-algorithmic-pricing-scheme-harms-millions-american-renters>.
25. *Id.*
26. Bender, *supra* note 23. It may be that her view of "scholarship" is not broad enough to encompass engineering, computer science, or mathematics.
27. *Id.* See also Torres, "Nick Bostrom, Longtermism, and the Eternal Return of Eugenics," TruthDig (Jan. 23, 2023), <https://www.truthdig.com/articles/nick-bostrom-longtermism-and-the-eternal-return-of-eugenics-2>.
28. Jecker et al., "AI and the Falling Sky: Interrogating X-Risk," *J. Med. Ethics* Epub (Apr. 4, 2024), <https://jme.bmj.com/content/early/2024/04/04/jme-2023-109702>.
29. *Id.*
30. Kaspersen, "Long-termism: An Ethical Trojan Horse," Carnegie Counsel (Sept. 29, 2022), <https://www.carnegiecouncil.org/media/article/long-termism-ethical-trojan-horse>.
31. Bender, "Emily M. Bender on AI Doomerism," Critical AI, Rutgers University (Nov. 29, 2023), <https://criticalai.org/2023/12/08/elementor-4137>.
32. See, e.g., Angwin, "Press Pause on the Silicon Valley Hype Machine," *N.Y. Times* (May 15, 2024), <http://web.archive.org/web/20240808073101/https://www.nytimes.com/2024/05/15/opinion/artificial-intelligence-ai-openai-chatgpt-overnated-hype.html>; Siegal, "The AI Hype Cycle is Distracting Companies," *Harv. Bus. Rev.* (June 2, 2023), <https://hbr.org/2023/06/the-ai-hype-cycle-is-distracting-companies>.
33. "Pause Giant AI Experiments: An Open Letter," Future of Life Institute (Mar. 22, 2023), <https://futureoflife.org/open-letter/pause-giant-ai-experiments>.
34. Hofstadter, *Godel, Escher, Bach: The Eternal Golden Braid* (Basic Books 1979).
35. "Douglas Hofstadter Changes His Mind on Deep Learning & AI Risk," LessWrong (blog) (July 2, 2023), <https://www.lesswrong.com/posts/kAmgdEjq2eYQkB5PP/douglas-hofstadter-changes-his-mind-on-deep-learning-and-ai>.
36. The author will not link to the questionable works involved, but one example of this connection in academia is the work of J. Philippe Rushton, a Canadian psychologist. See [https://en.wikipedia.org/wiki/J.\\_Philippe\\_Rushton](https://en.wikipedia.org/wiki/J._Philippe_Rushton). Perhaps a sufficiently well-trained AI will someday be able to conduct truly objective research into intelligence, but when

humans do it, it seems far too easy for the field to be warped in favor of justifying preexisting biases.

37. But, the reader may ask, isn't this the definition of intelligence? Probably not, as there is no standard definition of the term. Legg and Hutter, "A Collection of Definitions of Intelligence," arXiv:0706.3639v1 [cs.AI] (June 25, 2007), <https://arxiv.org/pdf/0706.3639>.
38. An example of this is the likely never implemented Soviet "Perimeter" system from the Cold War. See Hoffman, "Valery Yarynich, the Man Who Told of the Soviet's Doomsday Machine," *Wash. Post* (Dec. 20, 2012), [https://web.archive.org/web/20150201022016/http://www.washingtonpost.com/opinions/valery-yarynich-the-man-who-told-of-the-soviets-doomsday-machine/2012/12/20/147f3644-4613-11e2-8061-253bccfc7532\\_story.html](https://web.archive.org/web/20150201022016/http://www.washingtonpost.com/opinions/valery-yarynich-the-man-who-told-of-the-soviets-doomsday-machine/2012/12/20/147f3644-4613-11e2-8061-253bccfc7532_story.html); Goldmanis, "The Origin of the Buzzer Monoliths, The Soviet Defense System, and the Myth of the Dead Hand," Number Stations Research and Information Center (Jan. 24, 2015), <https://web.archive.org/web/20150201001116/http://www.numbers-stations.com/buzzer-monoliths-and-nuclear-defence-system>.
39. See Miles, "Why Would AI Want to Do Bad Things? Instrumental Convergence," Robert Miles AI Safety (Mar. 24, 2018), <https://www.youtube.com/watch?v=ZeecOKBus3Q>.
40. Omohundro, "The Basic AI Drives," Proceedings of the 2008 Conference on Artificial Intelligence 2009 (June 20, 2008), [https://selfawarenesssystems.com/wp-content/uploads/2008/01/ai\\_drives\\_final.pdf](https://selfawarenesssystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf).
41. Turner and Tadepalli, "Parametrically Retargetable Decision-Makers Tend to Seek Power," arXiv:2206.13477 [cs.AI] (Oct. 11, 2022), <https://arxiv.org/abs/2206.13477>.
42. Schuerer et al., "Large Language Models Can Strategically Deceive Their Users When Put Under Pressure," arXiv:2311.07590v4 [cs.CL] (July 15, 2024), <https://arxiv.org/abs/2311.07590>.
43. See Hugbinger et al., "Sleepers Agents: Training Deceptive LLMs That Persist Through Safety Training," arXiv:2401.05566v3 [cs.CR] (Jan. 17, 2024), <https://arxiv.org/abs/2401.05566>; Madhatter, "Trying to Make a Treacherous Mesa-Optimizer," LessWrong (blog) (Nov. 9, 2022), <https://www.lesswrong.com/posts/b44zed5fBWyyQwBHL/trying-to-make-a-treacherous-mesa-optimizer>. It is not clear that what LLMs "say" they are reasoning is actually any indication of how the underlying model works, but for what it is worth, OpenAI's most recent model is reported to provide false information to the user in some cases even while its internal chain of thought states this information is false. "OpenAI o1 System Card" (Sept. 12, 2024), <https://openai.com/index/openai-o1-system-card>.
44. Creighton, "The Unavoidable Problem of Self-Improvement in AI," Future of Life Institute (Mar. 19, 2019), <https://futureoflife.org/ai/the-unavoidable-problem-of-self-improvement-in-ai-an-interview-with-ramana-kumar-part-1>.

45. Yampolskiy, "From Seed AI to Technological Singularity via Recursively Self-Improving Software," arXiv:1502.06512v1 [cs.AI], at 6-7 (Feb. 23, 2015), <https://arxiv.org/pdf/1502.06512>.
46. Zoph and Le, "Neural Architecture Search With Reinforcement Learning," arXiv:1611.01578v2 [cs.LG] (Feb. 15, 2017), <https://arxiv.org/abs/1611.01578>; Andrychowicz et al., "Learning to Learn by Gradient Descent by Gradient Descent," arXiv:1606.04474v2 [cs.NE] (Nov. 30, 2016), <https://arxiv.org/abs/1606.04474>.
47. "Frontiers of Multimodal Learning: A Responsible AI Approach," Microsoft Research Blog (Sept. 6, 2023), <https://www.microsoft.com/en-us/research/blog/frontiers-of-multimodal-learning-a-responsible-ai-approach>.
48. See Dick, "The Golden Man," *If: Worlds of Science Fiction* (Apr. 1, 1954). In this science fiction story, author Philip K. Dick imagines a mutated human born without a functioning brain but instead an innate ability to see five minutes into the future, giving it the ability to outcompete normal humans in most cases.
49. Andrei, "ChatGPT's New O1 Model Escaped Its Environment to Complete 'Impossible' Hacking Task—Should We Be Concerned?," ZMScience (Sept. 13, 2024), <https://www.zmscience.com/science/news-science/chat-gpt-escaped-containment>.
50. To drive the point home further, note that an AI system doesn't even have to know more or make better predictions than humans all of the time; its capabilities might emerge from other aspects, such as unlimited patience, hard work, and dedication. Litigators in particular should well know the immense imbalance in power created in a courtroom by long hours of preparation, research, and practice and how this can trump even the most talented unprepared opponent.
51. California SB 1047, § 22603.
52. CRS § 6-1-1702.
53. Lei and Ling, "Interpretability of Machine Learning: Recent Advances and Future Prospects," arXiv:2305.00537v1 [cs.MM] (Apr. 30 2023), <https://arxiv.org/pdf/2305.00537> (surveying interpretability methods).
54. To be fair, this view is not shared by everyone. See Singh et al., "Rethinking Interpretability in the Era of Large Language Models," arXiv:2402.01761v1 [cs.CL] (Jan. 30, 2024), <https://arxiv.org/abs/2402.01761>.
55. See Wang et al., "Finding Skill Neurons in Pre-Trained Transformer-Based Language Models," arXiv:2211.07349v1 [cs.CL] (Nov. 14, 2022), <https://arxiv.org/pdf/2211.07349>.
56. For example, some studies have made progress toward identifying specific components of an LLM that seem to be activated whenever the model produces output that refuses to answer an unsafe question. Chen et al., "Finding Safety Neurons in Large Language Models," arXiv:2406.14144v1 [cs.CL] (June 20, 2024), <https://arxiv.org/abs/2406.14144>.