Law-Following AI

Law-Following AI 1: Sequence Introduction and Structure

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is written in my personal capacity, and does not necessarily represent the views of OpenAI or any other organization. Cross-posted to the Effective Altruism Forum.

This sequence of posts will argue that working to ensure that AI systems follow laws is a worthwhile way to improve the long-term future of AI.[1]

The structure of this sequence will be as follows:

- First, in this post, I will define some key terms and sketch what an ideal law-following AI ("LFAI") system might look like.
- In the next few posts, I will explain why law-following might not emerge by default given the existing constellation of alignment approaches, financial objectives, and legal constraints, and explain why this is troubling.
- Finally, I will propose some policy and technical routes to ameliorating these problems.

If the vision here excites you, and you would like to get funding to work on it, get in touch. I may be excited to recommend grants for people working on this, as long as it does not distract them from working on more important alignment issues.

Image by OpenAI's DALL·E.

Key Definitions

A **law-following AI** , or **LFAI** , is an AI system that is designed to rigorously comply with some defined set of human-originating rules ("laws"),[2] using legal interpretative techniques,[3] under the assumption that those laws apply to the AI in the same way that they would to a human. By "intrinsically motivated," I mean that the AI is motivated to obey those rules regardless of whether (a) its human principal wants it to obey the law,[4] or (b) disobeying the law would be instrumentally valuable.[5] (The Appendix to this post explores some possible conceptual issues with this definition of LFAI.)

I will compare LFAI with **intent-aligned AI**. The standard definition of "intent alignment" generally concerns only the relationship between some property of a human principal *H* and the actions of the human's AI agent *A*:

- Jan Leike et al. define the "agent alignment problem" as "How can we create agents that behave in accordance with the user's intentions?"
- Amanda Askell et al. define "alignment" as "the degree of overlap between the way two agents rank different outcomes."
- Paul Christiano defines "AI alignment" as "*A* is trying to do what *H* wants it to do."
- Richard Ngo endorses Christiano's definition.

Iason Gabriel does not directly define "intent alignment," but provides a taxonomy wherein an AI agent can be aligned with:

1. "Instructions: the agent does what I instruct it to do."
2. "Expressed intentions: the agent does what I intend it to do."
3. "Revealed preferences: the agent does what my behaviour reveals I prefer."
4. "Informed preferences or desires: the agent does what I would want it to do if I were rational and informed."
5. "Interest or well-being: the agent does what is in my interest, or what is best for me, objectively speaking."
6. "Values: the agent does what it morally ought to do, as defined by the individual or society."

All but (6) concern the relationship between *H* and *A*. It would therefore seem appropriate to describe them as types of intent alignment.

Alignment with some broader or more complete set of values—such as type (6) in Gabriel's taxonomy, Coherent Extrapolated Volition, or what Ngo calls "maximalist" or "ambitious" alignment—is perhaps desirable or even necessary, but seems harder than working on intent alignment.[6] Much current alignment work therefore focuses on intent alignment.

We can see that, on its face, intent alignment does not entail law-following. A key crux of this sequence, to be defended in subsequent posts, is that this gap between intent alignment and law-following is:

1. Bad in expectation for the long-term future.
2. Easier to bridge than the gap between intent alignment and deeper alignment with moral truth.
3. Therefore worth addressing.

To clarify, this sequence does **not** claim that LFAI can replace intent alignment.

A Sketch of LFAI

What might an LFAI system look like? I'm not a computer scientist, but here is roughly what I have in mind.

If *A* is an LFAI, then *A*'s evaluation of the legality of an action will sometimes trump *A*'s evaluation of an action in light of its benefit to *H*. In LFAI, as in a legally scrupulous human, legality constrains how an agent can advance their principal's interests. For example, a human mover may be instructed to efficiently move a box for her principal, but may not unnecessarily destroy others' property in doing so. Similarly, an LFAI

moving a box normally would not knock over a vase in its path, because doing so would violate the legal rights of the vase-owner.[7]

Above, I preliminarily defined LFAI as "rigorously comply[ing]" with some set of laws. Obviously this needs a bit more elaboration. We probably don't want to define this as *minimizing* legal noncompliance, since this would make the system extremely risk-averse to the point of being useless. More likely, one would attempt to weight legal downside risks heavily in the agent's objective function,[8] such that it would keep legal risk to an acceptable level.[9]

It is worth noting that LFAI is ideally not merely attempting to reduce its expected legal liability *in fact*. As will be explored later, a sufficiently smart agent could probably reduce its expected legal liability merely by hiding its knowledge/intentions/actions or corrupting a legal proceeding. An LFAI, by contrast, is attempting to obey the law in an idealized sense, even if it is unlikely to actually face legal consequences.

An LFAI system does not need to store all knowledge regarding the set of laws that it is trained to follow. More likely, the practical way to create such a system would be to make the system capable of recognizing when it faces sufficient legal uncertainty,[10] then seeking evaluation from a legal expert system ("Counselor").[11]

The Counselor could be a human lawyer, but in the long-run is probably most robust and efficient if (at least partially) automated. The Counselor would then render advice on the pure basis of idealized legality: the probability and expected legal downsides that would result from an idealized legal dispute regarding the action if everyone knew all the relevant facts.[12]

Thus pseudocode for an LFAI who wants to take an action *X* to benefit *H* might be:

1. If *X* is clearly illegal:
    1. don't do *X*.
2. Elseif *X* is maybe-illegal:
    1. Give Counselor all relevant information about *X* in an unbiased way; then
    2. Get Counselor's opinion on expected legal consequences from *X*; then
    3. Weigh expected legal consequences against benefit to *H* from *X*; then
    4. Decide whether to do *X* given those weightings.
3. Else:
    1. do *X*.

Note that this pseudocode may resemble the decisionmaking process of *A* if *H* wants *A* to obey the law. Thus, one route to giving an intent-aligned AI the motivation to obey the law may be stipulating to *A* that *H* wants *A* to obey the law.

With this picture in mind, it seems like, to make LFAI a reality, progress on the following open problems (non-exhaustively) would be useful:

- Reliably stipulating low-following conditions to AI systems' objectives.
    - Resolving any disagreement between law-following and a principal's instructions appropriately.
- Getting AI agents to recognize when they face legal uncertainty (especially in a way that does not incentivize ignorance of the law).
    - This seems similar to the intent alignment problem of getting agents to recognize when they need further information from principals, as in corrigibility work.

- Eliciting, in natural language, AI systems' honest description of its knowledge and desired actions.
    - As noted above, this seems likely to run into problems related to ELK generally.
- Mapping legal concepts of mental states (e.g., intent, knowledge) to features of AI systems.[13]
    - This seems related to interpretability and explainability work.
- Building Counselor functions.
    - Automating the process of legal research given a natural language description of an agent's proposed actions and mental state.
    - Simulating idealized and fair substantive legal disputes.
        - This seems related to Debate.

Appendix: More Conceptual Clarifications on LFAI

This Appendix provides some additional clarification on the definition of LFAI given above.

Applicability of Law to AI Systems

One might worry that the law often regulates physical behavior in a way that is not obviously applicable to all AI systems. For example, physical contact with another is an element of the tort of battery.[14] However, this may be less of a problem than initially appears: courts have been able to reason through whether to apply laws originating in meatspace to computational and cyberspace conduct.[15] Whether such analogies are *properly* applied is indeed highly debatable,[16] but the fact that such analogizing is conceptually possible reduces the force of this objection. Furthermore, if some laws are simply inapplicable to non-embodied actors, this is not a problem for the conceptual coherence of LFAI as a whole: an LFAI can simply ignore those laws,[17] and we can design laws specifically with computational content.

Perhaps a more fundamental problem is that the law frequently depends on *mental states* that are not straightforwardly applicable to AI systems. For example, the legality of an action may depend on whether the actor *intended* some harmful outcome. Thus, much of the value of LFAI depends on whether we can map human understandings of moral culpability to AI systems.

To me, however, this seems like an argument in favorof working on LFAI. Regardless of whether LFAI as such is valuable, if we expect increasingly autonomous AI systems to take increasingly impactful actions, we would probably like to understand how their objective functions (analogous to human *motives*) and world-model (analogous to human *knowledge*) map to their actions and the effects thereof. This is for the same reasons that we care about human motives and knowledge: when evaluating the alignment of agents, it is useful to know whether an agent intended to cause some harm, or knew that such a harm would ensue, etc. LFAI depends on progress on this, but is also potentially a useful toy problem for interpretability and related work in ML.

Predicting Legality

Legal compliance is also a function of both law and facts, and responsibility for definitive determinations of law and facts is split between judges and juries. Law often invokes standards like "reasonableness" that are definitively assessed only ex post, in the context of a particular dispute. The definitive legality of an action may therefore turn on an actual adjudication of the dispute. This is of course costly, which is why I suspect we would want an LFAI to act on its best estimate of what such an adjudication would yield (after asking a Counselor), rather than wait for such adjudication to take place.[18]

It is also worth distinguishing between whether an *actual court of law* would rule that an AI's behavior violated some law and whether a *simulated and fair legal dispute resolution process* (possibly including, for example, a bespoke arbitral panel) would conclude that the behavior violated the law. The latter may be more convenient for working on LFAI for a number of reasons, including that it can ignore or stipulate away some of the peculiarities of adjudicating disputes in which an AI system is a "party."

---

1. For early, informal discussion on this topic, see Michael St. Jules, *What are the challenges and problems with programming law-breaking constraints into AGI?*, **Effective Altruism Forum** (Feb. 2, 2020), https://forum.effectivealtruism.org/posts/qKXLpe7FNCdok3uvY/what-are-the-challenges-and-problems-with-programming-law [https://perma.cc/HJ4Y-XSSE] and accompanying comments. ↩

2. Whether such rules are actually encoded into legislation is not particularly important. Virtually all legal rules not part of public law can be made “legal” with regards to particular parties as part of a contract, for example. In any case, the heart of LFAI is being bound to follow rules, and interpreting those rules leveraging the rich body of useful rule-interpretation metarules from law. ↩

3. This is important because one of the core functions of law is to provide metarules regarding the interpretation of rules, guided by certain normative values (e.g., fairness, predictability, consistency). Indeed, rules of legal interpretation aim to solve many problems relevant to AI interpretation of instructions. *Cf.* Dylan Hadfield-Menell & Gillian Hadfield, *Incomplete Contracting and AI Alignment* (2018) (preprint), https://arxiv.org/abs/1804.04268. ↩

4. That is, the AI is not law-following *just because* the principal wants the AI to follow the law. Indeed, LFAI should disobey orders that would require it to behave illegally. ↩

5. That is, the AI is not law-following *just because* it is instrumentally valuable to it (because, e.g., being caught breaking the law would cause the AI to be turned off). ↩

6. As Ngo says, "My opinion is that defining alignment in maximalist terms is unhelpful, because it bundles together technical, ethical and political problems. While it may be the case that we need to make progress on all of these, assumptions about the latter two can significantly reduce clarity about technical issues." ↩

7. *Cf., e.g.*, Dario Amodei et al., Concrete Problems in AI Safety 4 (2016), https://arxiv.org/pdf/1606.06565.pdf. ↩

8. I don't here offer an opinion on what training regime would yield such an outcome—my hope is to get someone to answer that for me! ↩

9. This approach may work particularly well when combined with insurance requirements for people deploying AI systems. ↩

10. In the same way that an intent-aligned AI will sometimes ask for clarifications from a human principal. *See* Christiano. ↩

11. Note that there are ELK-style problems with this approach. If an AI is asking for legal advice and wants to minimize the negative signal it gets from the Counselor, it may hide certain relevant information (e.g., its true state of knowledge or its true intentions) from the Counselor. A good solution, as discussed, could be to simulate an idealized adjudication of the issue if all the parties knew all the relevant facts and had equal legal firepower. But incentivizing the LFAI to tell the Counselor its true knowledge/intentions is an ELK problem. In the limit, the Counselor need not strictly be a distinct agent from the LFAI: an LFAI system may have Counselor capabilities and run this "consultation" process internally. Nevertheless, it is illustratively useful to imagine a separation of the LFAI and the Counselor. ↩

12. This would be *idealized* so that details not ultimately relevant to the substantive legality of the action (e.g., jurisdiction, AI personhood, other procedural matters, asymmetries in legal firepower) can be ignored. See the final footnote of this piece for further discussion. ↩

13. See the Appendix for more discussion on this point. ↩

14. *See Battery*, **Wex** , https://www.law.cornell.edu/wex/battery (last accessed Sept. 3, 2021). ↩

15. *See, e.g.*, Intel Corp. v. Hamidi, 71 P.3d 296, 304–08 (Cal. 2003) (applying trespass to chattels to unauthorized electronic computer access); MAI Sys. Corp. v. Peak Computer, Inc., 991 F.2d 511, 518–19 (9th Cir. 1993) (storing data in RAM sufficient to create a "copy" for copyright purposes, despite the fact that a "copy" must be "fixed in a tangible medium"); *cf.* United States v. Jones, 565 U.S. 400, 406 n.3 (2012) (analogizing GPS tracking to in-person surveillance for Fourth Amendment purposes). ↩

16. *See, e.g.*, Jonathan H. Blavin & I. Glenn Cohen, *Gore, Gibson, and Goldsmith: The Evolution of Internet Metaphors in Law and Commentary*, 16 **Harv. J.L. & Tech.** 265 (2002). ↩

17. However, the case for working on LFAI certainly diminishes with the number of applicable laws. ↩

18. This raises further issues, including the possibility of self-reference. For example, an LFAI or Counselor asymmetrically deployed by one litigant may be able to persuade a judge or jury of its position, even if it's not the best outcome. To avoid this, such simulations should assume that judges and juries are fully apprised of all relevant facts (i.e., neither the LFAI nor Counselor can obscure relevant evidence) and if deployed in the simulated proceeding are symmetrically available to both sides. ↩

Law-Following AI 2: Intent Alignment + Superintelligence → Lawless AI (By Default)

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is written in my personal capacity, and does not necessarily represent the views of OpenAI or any other organization. Cross-posted to the Effective Altruism Forum.

In the first post of this sequence, I defined "law-following AI" ("LFAI") and "intent alignment." In this post, I will begin to motivate the case for working on LFAI by showing that an AI agent *A* that is only intent-aligned (i.e., without additional law-following constraints directly on the agents) would in many cases break the law to advance its human principal *H*'s interests.

If you agree that it is generally good for agents to be bound by law, then this should be a pro tanto reason to support work on LFAI in addition to intent alignment. If you need more convincing that LFAI would be generally good, the next few posts will explore some more specific motivations for LFAI.

Image by OpenAI's DALL·E

Sufficiently Intelligent Agents Can Evade Detection and Attribution

I suspect this point will need little argumentation to many of my readers. Evasion and obfuscation seems likely to be a generally promising means of ensuring self-preservation (because detection would increase the probability of punishment and constraint), which in turn is plausibly a convergent instrumental goal of optimizing AI systems.[1]

For example, to evade detection and attribution, *A* could:

- Behave innocently only when being observed;
- Orchestrate complex malevolent actions through a coordinated series of small, inscrutable or innocent-seeming steps;
- Surreptitiously create and run subagents or copies of itself on new hardware, potentially very far removed from its original hardware;
- Convince humans or other systems to take actions on its behalf, without revealing its true identity; and/or
- Delete logs of its actions.

A sufficiently intelligent AI agent could likely consistently fool humans using these and other (unforeseen) techniques.

Other Ways to Circumvent Law

Even in the best case scenario, where the agent is detected and within the jurisdiction of a well-functioning legal system, it would be reasonable to question whether *A* or *H* could be effectively subject to normal legal processes.[2] If *A* had a motivation to, *A* could help *H* escape liability by, for example:[3]

- "Outlawyering" counterparties.
- Benefitting *H* in a way that would undermine recourse for creditors.
- Shifting and hiding assets in ways that would make it difficult for creditors to reach.[4]
- Persuasively arguing for the law to be changed in *H*'s favor (by legislation or otherwise).
- Engaging in vexatious litigation techniques to delay and raise the costs of the proceeding.
- Convincingly fabricating favorable evidence and destroying or obscuring unfavorable evidence.
- Bribing, misleading, or intimidating counterparties, witnesses, jurors, and judges.

A Competent Intent-Aligned Agent Will Sometimes Intentionally Break the Law

As I said in the previous post, on its face, intent-alignment does not entail law-following. Part of law is coercing prosocial behavior:[5] law incentivizes agents to behave in ways that they do not intrinsically want to behave. If *A* is aligned with *H*, whether *A* obeys the law depends on whether *H* **wants** *A* to obey the law. Subsequent posts will examine what legal consequences *H* might face if *A* causes legally cognizable harms. However, even if an adequate theory of liability for the *H* was available, it will seem impossible to hold *H* liable if nobody can produce evidence that some agent of *H*'s was responsible for those harms. As argued above, a sufficiently intelligent agent probably *could* consistently avoid leaving any such evidence.

Detection and attribution would not solve the problem, however. Even if *H* was compelled, under court order, to instruct *A* to behave in some way, it's not clear that *A* would follow the order. Consider again Iason Gabriel's taxonomy of alignment. We can see that, for most types of intent alignment, an intent-aligned agent would likely not obey compelled instructions that are against *H*'s true wishes:

1. "Instructions: the agent does what I instruct it to do."

    - If *H* asks the agent to stop, it will. However, this type of alignment is likely to be insufficient for safety.[6]

2. "Expressed intentions: the agent does what I intend it to do."
    - Even if *H* coerced to instruct the agent to stop, a sufficiently intelligent agent will probably not follow the instruction. Even though *H* has instructed *A* to stop, by supposition, *H* does not actually want *A* to stop. Under most definitions of intent alignment,[7] it would therefore not follow the order.[8]

3. "Revealed preferences: the agent does what my behaviour reveals I prefer."
    - *H*'s revealed preference would probably be that *A* not follow the order.[9] Therefore, *A* would not obey the order.

4. "Informed preferences or desires: the agent does what I would want it to do if I were rational and informed."
    - *H*'s rational and informed preference would probably be that *A* not follow the order. Therefore, *A* would not obey the order.

5. "Interest or well-being: the agent does what is in my interest, or what is best for me, objectively speaking."
    - It is in *H*'s objective best interest for *A* to disobey the order. Therefore, *A* would not obey the order.

Now, it may be the case that *H* actually *does* want *A* to obey the order, though compelled, if the failure of *A* to obey would lead to liability for *H* that is worse than the results of *A*'s obedience (e.g., because *H* will be held in contempt of court if *A* does not actually obey). However, note that "[o]rdinarily, one charged with contempt of court for failure to comply with a court order makes a complete defense by proving that he is unable to comply."[10] *H* can comply with an order that requires *H* to **command** *A* to do something, but it may be impossible for *H* to actually **force** *A* to comply if the order is against *H*'s true wishes (to which *A* is aligned).[11] If so, *H* could have an impossibility defense to contempt.[12] *A*, understanding this, may continue on without complying because *A* understands that *H* will not actually be held in contempt. *H* can therefore benefit from *A*'s disobedience. *A* will therefore be lawless.

Appendix: The Impossibility Defense

A's behavior here would be functionally similar to a trustee acting pursuant to a distress clauses in asset protection trusts ("APTs").[13] While these provisions can prevent a contempt charge, the burden of proof on the alleged contemnor is high. [14]

As a matter of policy, however, courts may decide to pre-commit to a contempt standard that does not allow for an impossibility defense when the defendant's AI agent refuses to obey orders issued pursuant to a court order. Analogously, courts are imposing heightened impossibility standards in response to APTs, in an attempt to make their use more onerous.[15] If this pre-commitment is credible, it may change the agent's behavior because *H* may then genuinely desire *A* to perform (because *H* will be held in contempt otherwise). However, such a policy may be both contrary to precedent and more fundamental notions of fairness and due process: in some cases

A's refusal to comply may be a surprise to *H*, since *H* may have had a long history of observing *A* scrupulously complying with *H*'s orders, and *H* did not implement principal–agent alignment for the purpose of evading court orders. If so, *H* may be able to invoke impossibility more easily, since the impossibility was not as clearly intentionally self-induced as in the APT case. Furthermore, I would intuitively not expect courts to advance such a reform until they have faced multiple such instances of AI disobedience. This seems bad if we expect the earliest deployed AI agents to have an outsized impact on society. In any case, I would expect the *possibility* of favorable law reform post-AGI to solve this problem to be an insufficient solution. Finally, I would expect sufficiently intelligent agents to recognize these dynamics, and attempt to find ways to circumvent the contempt process itself, such as by surreptitious non-compliance.

An alternative, pre-AGI solution (which arguably seems pretty sensible from a public policy perspective anyway) is to advocate weakening the impossibility defense for self-imposed impossibility.

---

1. *See generally* Alexander Matt Turner et al., Optimal Policies Tend To Seek Power (version 9, 2021) (preprint), https://arxiv.org/abs/1912.01683. ↩

2. Even this may not hold for many types of agreements, including in particular international treaties. ↩

3. *See also* **Cullen O'Keefe et al., The Windfall Clause: Distributing the Benefits of AI for the Common Good** 26–27 (2020), https://perma.cc/8KES-GTBN; Jan Leike, On The Windfall Clause (2020) (unpublished manuscript), https://docs.google.com/document/d/1leOVJkNDDj-NZUZrNJauZw9S8pBpuPAJotD0gpnGEig/. ↩

4. Indeed, this is already a common technique without the use of AI systems. ↩

5. "If men were angels, no government would be necessary." **The Federalist** No. 51. This surely overstates the point: law can also help solve coordination problems and facilitate mutually desired outcomes. But prosocial coercion is nevertheless an important function of law and government. ↩

6. *See* Gabriel at 7 ("However, as Russell has pointed out, the tendency towards excessive literalism poses significant challenges for AI and the principal who directs it, with the story of King Midas serving as a cautionary tale. In this fabled scenario, the protagonist gets precisely what he asks for—that everything he touches turns to gold—not what he really wanted. Yet, avoiding such outcomes can be extremely hard in practice. In the context of a computer game called CoastRunners, an artificial agent that had been trained to maximise its score looped around and around in circles ad infinitum, achieving a high score without ever finishing the race, which is what it was really meant to do. On a larger scale, it is difficult to precisely specify a broad objective that captures everything we care about, so in practice the agent will probably optimise for some *proxy* that is not completely aligned with our goal. Even if this proxy objective is 'almost' right, its optimum could be disastrous according to our true objective." (citations omitted)). ↩

7. Based on my informal survey of alignment researchers at OpenAI. Everyone I asked agreed that an intent-aligned agent would not follow an order that the

principal did not actually want followed. *Cf. also*Christiano (A is aligned when it "is trying to do what H *wants* it to do" (emphasis added)). ↩

8. We can compare this definition of intent with to the relevant legal definition thereof: "To have in mind a fixed purpose to reach a desired objective; to have as one's purpose." *INTEND*, **Black's Law Dictionary** (11th ed. 2019). *H* does not "intend" for the order to be followed under this definition: the "desired objective" of *H* issuing the order is to follow *H*'s legal obligations, not actually achieve the result contemplated by the order. ↩

9. For example, *H* would exhibit signs of happiness when *A* continues. ↩

10. United States v. Bryan, 339 U.S. 323, 330 (1950). ↩

11. A principal may want its AI agents to be able to distinguish between genuine and coerced instructions, and to disobey the latter. Indeed, this might generally be a good thing, except for the case when compulsion is pursuant to law rather than extortion. ↩

12. See Appendix for further discussion. ↩

13. *See generally Asset Protection Trust*, **Wex** , https://www.law.cornell.edu/wex/asset_protection_trust (last visited Mar. 24, 2022); Richard C. Ausness, *The Offshore Asset Protection Trust: A Prudent Financial Planning Device or the Last Refuge of A Scoundrel?*, 45 **Duq. L. Rev.** 147, 174 (2007). ↩

14. *See generally* 2 **Asset Protection: Dom. & Int'l L. & Tactics** §§ 26:5–6 (2021). ↩

15. *See id.* ↩

Law-Following AI 3: Lawless AI Agents Undermine Stabilizing Agreements

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is written in my personal capacity, and does not necessarily represent the views of OpenAI or any other organization. Cross-posted to the Effective Altruism Forum

In the previous post of this sequence, I argued that intent-aligned AIs would, by default, have incentives to break the law. This post goes into one particularly bad consequence of that incentive: the increased difficulty of making credible pre-AGI commitments about post-AGI actions.

Image by OpenAI's DALL·E

In AGI policy and strategy, it would often be useful to adopt credible commitments about what various actors will do post-AGI. For example, it may be desirable for two leading nations in AGI to agree to refrain from racing to build AGI (at the potential cost to AGI safety) and instead split the economic upside from AGI, thereby transforming a negative-sum dynamic into a positive-sum one.[1] Nations might agree to forego their own development of militarily useful AI systems if they receive security assurances from states that *do* choose to pursue such systems,[2] thus reducing the number of actors pursuing potentially dangerous military AI development, and therefore reducing the risk of a catastrophic accident. One can imagine similar types of pre-AGI agreements concerning space colonization post-AGI, the welfare of digital minds, democratic control of AGI, and many more important issues in the post-AGI world.

Lawless AI poses an enormous problem for such agreements. Agreements between entities, such as states or companies, require performance by agents of the parties: a state or corporation only acts through its agents. But if one party cannot convincingly show that their agents will carry out orders to comply with such agreements, the other party will be very reluctant to enter into an agreement with them. All the more so if the principal cannot prevent its agents from working *against* aspects of the agreement that are detrimental to the principal.

I will use the following as a toy problem to show what I mean more concretely: *X* and *Y* are two leading AGI development firms considering agreeing to split the profits from AGI development if they both agree to common minimum implement safety measures to prevent racing.[3] How might this play out?

X might worry that *Y* will develop an AGI agent, *A(Y)*, with the objective of maximizing the value of *Y*'s shares. *A(Y)* can do this in ways that would undermine *X*'s recourse to *Y*'s assets. For example, *A(Y)* could create a subagent that surreptitiously earns money and directly mails checks to the shareholders of *Y*, as a sort of pseudo-dividend. That money might never pass through *Y*, so *X* would have a hard time reaching it, even if they detected this. *X* might also worry that *A(Y)* would also corrupt the legal process in *Y*'s favor, so that even if they had a good legal claim, they could not vindicate it in court at any reasonable cost.

With these concerns in mind, the deal is unlikely to happen.

How might LFAI improve the situation? *X* and *Y* could both agree that:

1. All sufficiently advanced AI agents ("Qualifying Agents") that they train, use, benefit from, or deploy will be LFAI.
2. The other party may, through a neutral third-party bound by confidentiality,[4] inspect Qualifying Agents and their development to verify that the trained agent will be LFAI when deployed.
3. Qualifying Agents will be specifically instructed to assume that they are bound by the agreement.[5]
4. The parties will generally cause all of their agents to give full effect to the agreement.
5. In any proceeding where the parties are adverse, neither party shall use any artificial intelligence technology to its legal or strategic benefit, unless the other party has free and equal access to the same technology.[6]
6. In case of any ambiguity in the terms of the agreement, the ambiguity shall be resolved in favor of more constraints on Qualifying Agents and their principal(s).

The resulting agreement would probably not be watertight, but the parties would have much greater assurances of post-AGI performance than they would absent LFAI. Referring back to the example above, *A(Y)* could not surreptitiously earn money outside of *Y* and pass it on the *Y*'s shareholders, nor could *A(Y)* illegally corrupt adjudications of the agreement. LFAI, then, could make the parties much more likely to be willing to enter stabilizing pre-AGI agreements that require post-AGI performance. This seems like a substantial improvement.

---

1. *Cf.* Amanda Askell et al., The Role of Cooperation in Responsible AI Development (2019) (preprint), https://arxiv.org/abs/1907.04534. ↩

2. Of course, this could be analogized to similar agreements regarding nuclear disarmament, such as Ukraine's fateful decision to surrender its post-Soviet nuclear arsenal in exchange for security assurances (which have since been violated by Russia). *See, e.g.*, Editorial, *How Ukraine Was Betrayed in Budapest*, **Wall St. J.** (Feb. 23, 2022), https://www.wsj.com/articles/how-ukraine-was-betrayed-in-budapest-russia-vladimir-putin-us-uk-volodymyr-zelensky-nuclear-weapons-11645657263?reflink=desktopwebshare_permalink. Observers (especially those facing potential conflict with Russia) might reasonably question whether any such disarmament agreements are credible. ↩

3. We will ignore antitrust considerations regarding such an agreement for the sake of illustration. ↩

4. So that this inspection process cannot be used for industrial espionage. ↩

5. This may not be the case as a matter of background contract and agency law, and so should be stipulated. ↩

6. This is designed to guard against the case where one party develops AI super-lawyers, then wields them asymmetrically to their advantage. ↩

Law-Following AI 4: Don't Rely on Vicarious Liability

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is written in my personal capacity, and does not necessarily represent the views of OpenAI or any other organization. Cross-posted to the Effective Altruism Forum.

Image by OpenAI's DALL·E

If an agent *A* causes some harm while intending to benefit a principal *P*, what is *P*'s liability? The answer to this question is important because any liability to *P* should affect *A*'s calculus (insofar as *A* is trying to benefit and avoid harming *P*). Liability to *P* would help deter *A* from causing harm.

If *A* is a human, the law currently provides at least two mechanisms for discouraging :

1. Making *A* directly liable, and
2. Making *P* vicariously liable for *A*'s actions.[1]

What if *A* is an AI? AIs are not (yet?) legal persons, and so cannot yet be held directly liable. Thus, the main legal deterrent would have to work on *P* or some other person in the causal chain, such as the developer of the AI agent (who may not be the principal). However, there are several reasons to worry about relying on this as a strategy to make *A* compliant (which, in this case, means not tortiously harming others) under the current state of law and AI.

First, the problem of evasion still remains. Because sufficiently intelligent agents can evade detection and attribution, *A* may often (perhaps usually) prefer evasion over compliance when compliance would hinder *A*'s ability to benefit *P*.

Second, the applicability and appropriateness of various theories of vicarious liability to the actions of AI agents is heavily debated in legal scholarship.[2] These debates have cast some doubt on whether/which harms from AI "agents" can properly give rise to liability to human principals.[3] Other possible theories of human liability—such as products liability—also face doctrinal challenges.[4]

Third, note that relying on vicarious liability alone leaves *A* under fewer constraints than an analogous human would. Under most vicarious liability regimes, *A* would still be directly liable for her actions, even if *P* would also be vicariously liable. It seems unwise to legally constrain *A less* than we constrain humans in analogous circumstances.

Finally (and most decisively in my opinion) developing a theory that assigns liability to *P* based on *A*'s actions (or actions + "mental" state) dramatically lowers the bar for creating an LFAI system in the first place. Once we have such a theory, *A* (if intent-aligned) should indeed incorporate expected vicarious liability to *P* into its decision procedure. However, if *A* is already reasoning about whether its actions would violate law (as required to make vicarious liability an effective constraint on *A*'s actions), **it seems strictly better to require *A* to directly incorporate that information into its decision procedure**, rather than needing to go through the additional step of estimating the expected liability to *P*. This direct approach is simpler, and removes the possibility of evasion as a way around the legal constraint.[5]

Further legal scholarship on vicarious liability for AI systems may still be valuable. If most morally significant autonomous activity in the world is indeed carried out by AI agents in the future, incentivizing their principals to constrain them seems important. But I think there are good reasons to suppose that this will either be ineffective for—or else dominated by—requiring AIs to be directly law-following.

---

1. *See generally Vicarious Liability*, **Wex**, https://www.law.cornell.edu/wex/vicarious_liability (last accessed Sept. 10, 2021);

Respondeat Superior, **Wex**, https://www.law.cornell.edu/wex/respondeat_superior (last accessed Sept. 10, 2021). ↩

2. *See, e.g.*, Benny Chan, *Applying A Common Enterprise Theory of Liability to Clinical AI Systems*, 47 **Am. J.L. & Med.** 351 (2021); Mihailis E. Diamantis, *Algorithms Acting Badly: A Solution from Corporate Law*, 89 **Geo. Wash. L. Rev.** 801 (2021); Mihailis E. Diamantis, *The Extended Corporate Mind: When Corporations Use AI to Break the Law*, 98 **N.C. L. Rev.** 893 (2020); Yaniv Benhamou & Justine Ferland, *Artificial Intelligence & Damages: Assessing Liability and Calculating the Damages*, *in* **Leading Legal Disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law** (forthcoming 2020), https://ssrn.com/abstract=3535387; Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 **U. Chi. L. Rev.** 1311 (2019); Elizabeth Fuzaylova, *War Torts, Autonomous Weapon Systems, and Liability: Why A Limited Strict Liability Tort Regime Should Be Implemented*, 40 **Cardozo L. Rev.** 1327 (2019); Bryan H. Choi, *Crashworthy Code*, 94 **Wash. L. Rev.** 39 (2019); Xavier Frank, *Is Watson for Oncology Per Se Unreasonably Dangerous?: Making A Case for How to Prove Products Liability Based on A Flawed Artificial Intelligence Design*, 45 **Am. J.L. & Med.** 273 (2019); Matthew U. Scherer, *Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems*, 19 **Nev. L.J.** 259 (2018); David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 **Wash. L. Rev.** 117, 121–124 (2014); Jessica S. Allain, *From* Jeopardy! *To Jaundice: The Medical Liability Implications of Dr. Watson and Other Artificial Intelligence Systems*, 73 **La. L. Rev.** 1049 (2013). ↩

3. *See* Benhamou & Ferland, *supra*, at 13; Vladeck, *supra*, at 123 n.21. ↩

4. Diamantis, *Algorithms Acting Badly*, *supra*, at 823–26 (arguing that products liability will largely be unavailable); Vladeck, *supra*, at 129–41; Scherer, *supra*, at 280–81. ↩

5. Specifically, in the direct case, what matters is whether *A* is actually violating the law, whereas in the vicarious case, what matters is the *expected liability* to *P*. *A* can reduce expected liability to P through evasion, but cannot reduce the probability of "actually" breaking the law through evasion. ↩