[Draft-- 06-01-2024]

The Law of AI is the Law of Risky Agents without Intentions

Ian Ayres[1] and Jack M. Balkin[2]

I. Introduction

A recurrent problem in adapting law to artificial intelligence programs is how the law should regulate the use of entitles that lack intentions. Many areas of the law, including freedom of speech, copyright, and criminal law, make liability turn on whether the actor who causes harm (or creates a risk of harm) has a certain intention or *mens rea.* But AI agents—at least the ones we currently have—do not have intentions in the way that humans do. If liability turns on intention, that might immunize the use of AI programs from liability.

We think that the best solution is to employ objective standards that are familiar in many different parts of the law. These legal standards either ascribe intention to actors or hold them to objective standards of conduct.

Of course, the AI programs themselves are not the responsible actors; instead, they are technologies used by human beings that have effects on other human beings. Therefore, the real question of legal obligation is who should be held responsible for the use of AI and under what conditions.[3]

We might think of AI programs as acting on behalf of human beings. Then AI programs are like agents that lack intentions but that create risks of harm to people. It follows that the law of AI is the law of risky agents without intentions. The law should hold these risky agents to objective standards of behavior. Holding AI agents to objective standards of behavior, in turn, means holding the people and organizations that implement these technologies to standards of reasonable care and requirements of reasonable reduction of risk. Thus, if the law of AI is the law of risky agents without intentions, then legal regulation of AI must require AI companies to internalize the costs of the risks they impose on society, through rules and standards that regulate design, training, and implementation.[4] To regulate AI, one must regulate the risks created by the people and organizations that employ AI, including their choices to use AI in the first place.

---

[1] Oscar M. Ruebhausen Professor, Yale Law School.

[2] Knight Professor of Constitutional Law and the First Amendment, Yale Law School. Harran Deu provided helpful research assistance.

[3] Jack M. Balkin, *2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data*, 78 Ohio St. L.J. 1217, 1223 (2017) (“the problem isn't the robots; it's the humans”).

[4] See Margot Kaminski, *Regulating the Risks of AI*, 103 B.U. L. Rev. 1347, 1351 (2023)("lawmakers in both the United States and the European Union (“EU”) have turned to the tools of risk regulation to govern AI systems.")

The law holds principals liable for their agents in several ways. First, principals are liable for what agents do on their behalf, for example when agents enter into contracts on behalf of their principals.[5] The doctrine of respondeat superior in tort law is a special case of this idea.

When people use AI, they should be responsible for the harms that occur when the risks of using the technology are realized. This is similar to the idea that principals are responsible for what their agents do on their behalf. The legal responsibility of the AI agents is then ascribed to human beings who stand in the position of the principal. Because AI agents lack intentions, the law should hold them (and the people and the companies that employ them) to objective standards, which might be negligence, strict liability or the highest level of care (when the AI agent performs the functions of a fiduciary). In general, people should not be able to obtain a reduced duty of care by substituting an AI agent for a human agent. Thus, if an AI agent makes a contract on behalf of a principal, the principal may not object that they did not know what the AI agent was doing; if the AI performs fiduciary services, the principal must be held to the highest standard of care, and so on.

Second, principals can be liable for failing to exercise due care in supervising or training their agents when their agents cause harms to another. In the same way, human beings may be held liable for negligently employing an AI program that is not properly trained and regulated and that causes harm to others. People may also be held responsible for designing, programming, or training a risky AI that causes harm to others. The best analogy here is to liability for defective design, which employs a risk-utility test.[6]

Third, and finally, human beings may be liable for hosting an AI program for others to use that creates foreseeable risks of harm to the users or to third parties. Here too, liability should turn on objective standards of reasonableness and risk reduction. If a reasonable person would know that the AI program may cause harm, the person hosting the program has a duty to make the program reasonably safe to use in ordinary circumstances.

In each case, the point of using objective standards for the performance of AI programs is to make it easier to assign responsibility to human beings, who are always the real parties in interest. Technology is not a relationship between human beings and things. It is a social relationship between different groups of human beings that is mediated by the deployment and use of technologies.[7] In the most basic sense the question of robotics and AI regulation is the question of what the law should do when human beings (and human-run companies and corporations) implement new technologies that substitute for human thought and action and have

[5] See. e.g., CT Uniform Electronic Transaction Act Sec. 1-279 (1)("A contract may be formed by the interaction of electronic agents of the parties, even if no individual was aware of or reviewed the electronic agents' actions or the resulting terms and agreements."). Contract law has for decades contemplated that contracts can be created by electronic agents. See Anthony J. Bellia, Jr., *Contracting with Electronic Agents*, 50 Emory L. J. 1047 (2001).

[6] Restatement (Third) of Torts, Product Liability (1997 ed.). §2(b) ("a product is defective when, at the time of sale or distribution, it contains a manufacturing defect, is defective in design or is defective because of inadequate instructions or warnings.").

[7] Balkin, *The Three Laws of Robotics in the Age of Big Data*, *supra* note 3 ("These technologies mediate social relations between human beings and other human beings. Technology is embedded into--and often disguises--social relations.").

effects on other human beings.[8] These are the people who are the real parties in interest in the deployment of technology and who should bear responsibility for its use.

II. Ascribed intentions and objective standards

The law has long dealt with the problem that it is often difficult to know what other people are thinking. And in regulating partnerships, associations, and artificial persons like corporations, the law often has to deal with entities that either lack a single human intention or lack intentions altogether.

The law has two basic strategies for these situations that we should adapt to artificial intelligence programs. The first strategy is that the law sometimes *ascribes intention* to an entity. For example, in intentional tort the law ascribes intentions to actors. Actors who injure others are presumed to have intended the consequences of their actions.[9] The second strategy *holds actors to a standard of behavior*—usually one of reasonableness— regardless of their actual intentions. This standard of behavior is external to the actors' real intentions. Thus, in negligence law, one is liable if one behaves in a way that a reasonable person would not behave. All persons, regardless of their actual knowledge and intentions, must live up to the standard of reasonable care.[10] In contract law, the intention of a party is measured according to an objective standard—what a reasonable person would interpret the party's intentions to have been.[11] In agency law, an agency relationship exists according to the judgment of a reasonable person.[12] The standard of behavior need not always be reasonableness. In fiduciary law, the fiduciary is held to the highest standard of care in the interests of the client or principal.[13]

The two strategies of ascribing intention and imposing standards of behavior based on an imagined intention are mirror images of each other. The first strategy says "regardless of your intentions, the law will treat you as if you had a particular intention and regulate or penalize you accordingly." The second strategy says "regardless of your actual intentions, the law will measure your conduct by the standard of a hypothetical person with a particular mental state and regulate or penalize you if you do not live up to that standard."

We propose that the law regulate the use of AI programs through these two strategies. First, where necessary, the law should ascribe intentions to AI programs. The law should presume that

---

[8] See *id.* at 1224 (describing the "substitution effect" of AI and robotics).

[9] *See* Restatement (Second) of Torts §8A, comment b ("If the actor knows that the consequences are certain, or substantially certain, to result from his act, and still goes ahead, he is treated by the law as if he had in fact desired to produce the result.")

[10] Restatement (Third) of Torts §7(a) ("An actor ordinarily has a duty to exercise reasonable care when the actor's conduct creates a risk of physical harm.").

[11] See Restatement (Second) of Contracts, § 212 Comment a ("the relevant intention of a party is that manifested by him rather than any different undisclosed intention").

[12] Restatement (Third) of Agency §2.02 (3) (AM. L. INST. 2006), ("An agent's understanding of the principal's objectives is reasonable if it accords with the principal's manifestations and the inferences that a reasonable person in the agent's position would draw from the circumstances creating the agency.")

[13] *SEC v. Capital Gains Research Bureau, Inc.*, 375 U.S. 180, 194 (1963).

AI programs intend the reasonable and foreseeable consequences of their actions. Second, where the law requires a particular mens rea to hold a human being (or corporation) liable, the appropriate standard for liability for use of AI programs should be based on an objective standard. Where the law requires a showing of recklessness, knowledge, or purpose for a human actor, the behavior of AI programs should be regulated by a requirement of reasonableness on the part of those who design, maintain, and implement the programs. Where the law applies a higher standard—for example, negligence, highest care (as in fiduciary situations), or strict liability—the rule should be the same for AI programs as it is for humans. Thus, a company that employs AI programs to perform fiduciary functions must offer the highest care to its clients and beneficiaries and train, regulate, and maintain the program accordingly.[14]

III. Why ascribed intentions and objective standards?

When the law imposes less than an objective standard on human beings, the best justifications concern the preservation and protection of human liberty, and the fear that objective standards will chill innocent (and even valuable) action as well as morally culpable conduct.

Take First Amendment law as an example. If a person defames a public figure, the plaintiff must show actual malice—knowledge that the published statement was false or made with reckless disregard for its falsity.[15] The best justification of this high standard is that otherwise people would be chilled from making valuable contributions to public discourse. It is true that some of this speech will not improve the quality of public discussion; nevertheless, the law strikes a balance in favor of preserving the liberty to speak over the harm to reputation.

Recently, in *Counterman v. Colorado*,[16] the Supreme Court made clear that when a person is criminally prosecuted for making online threats, the state must show that the defendant subjectively understood that they were putting other people in fear of their safety.[17] As Justice Kagan explained, "[p]rohibitions on speech have the potential to chill, or deter, speech outside their boundaries. A speaker may be unsure about the side of a line on which his speech falls. Or he may worry that the legal system will err, and count speech that is permissible as instead not. Or he may simply be concerned about the expense of becoming entangled in the legal system. The result is "self-censorship" of speech that could not be proscribed.... And an important tool to prevent that outcome …is to condition liability on the State's showing of a culpable mental state."[18] To be sure, "[s]uch a requirement comes at a cost: It will shield some otherwise proscribable … speech …. But the added element reduces the prospect of chilling fully protected

[14] Cf. Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. Davis L. Rev. 1183 (2016) (arguing that digital companies have fiduciary obligations toward those whose data they collect and use.)
[15] *See e.g.* The New York Times Co. v. Sullivan 376 U.S. 254, 279–80 (1964).
[16] Counterman v. Colorado, 600 U.S. 66 (2023).
[17] *Id.* at 73.
[18] *Id.* at 75.

expression."[19] The Court chose a subjective standard to protect human liberty and the ability to participate in public discussion.

These arguments for subjective standards do not apply to artificial intelligence programs. The programs that we have today exercise neither individual autonomy nor political liberty. Moreover, AI is not chilled from expressing itself in the way that Justice Kagan describes. Therefore, the interest in preserving and protecting human liberty that justifies the use of subjective standards does not apply, and cannot outweigh the harms to society.

To be sure, First Amendment law protects not only the rights of speakers but also the rights of listeners to receive ideas and information.[20] And sometimes, as in commercial speech, First Amendment doctrine is based on the right of listeners to receive information, not speakers' rights to provide it.[21] In general, however, the standard of First Amendment protection is the same whether we view speech in terms of speaker or listener rights. For example, if a speaker defames a public figure with actual malice, we do not insulate the speaker from liability because listeners might also want to hear the speech. So even if the basis for protecting AI speech is the rights of listeners, it would not justify applying a subjective standard to AI programs, particularly given that AI programs have no subjective intentions.

IV. Applications

We offer two applications of our thesis: defamation and copyright infringement. In both cases we argue that the appropriate way to solve questions of liability involves objective standards and reasonable regulation of the risks created by AI technologies.

A. Defamatory Hallucinations

Large language models often state things that are not true, especially when prompted to do so. Suppose that *A* asks a LLM to list all of the crimes committed by *B*, and the model responds with a list of non-existent crimes and the dates they happened.[22] The LLM does not have intentions, so it makes no sense to ask whether it acted with actual malice. Even where the plaintiff is a

---

[19] *Id.*

[20] *See* Toni M. Massaro, Helen Norton, and Margot E. Kaminski, *Siri-ously 2.0: What Artificial Intelligence Reveals About the First Amendment*, 101 Minn. L. Rev. 2481, 2482 (2017)(arguing that defenses of constitutional rights for the use of AI will be premised on listeners' rights.)

[21] Robert Post & Amanda Shanor, *Adam Smith's First Amendment*, 128 Harv. L. Rev. F. 165, 170 (2015) (explaining that while "[o]rdinary First Amendment doctrine . . . focuses on the rights of speakers, not listeners," the "constitutional value of commercial speech lies in the rights of *listeners* to receive information so that they might make intelligent and informed decisions").

[22] OpenAI has recently been sued for incorrectly summarizing a complaint saying that Mark Walters was accused of defrauding and embezzling funds from the Second Amendment Foundation. https://aboutblaw.com/8ts. See Isaiah Porter, *OpenAI Hit With First Defamation Suit Over ChatGPT Hallucination*, Bloomberg Law, https://news.bloomberglaw.com/tech-and-telecom-law/openai-hit-with-first-defamation-suit-over-chatgpt-hallucination

private party, and a negligence standard applies, one should not analogize the LLM to a journalist or author. The law expects human beings to research a topic and investigate sources of information before publication. But LLMs do not work this way. Rather, an LLM is a technology designed by a party (or some combination of parties) to respond to end-user prompts. It is a prediction model that generates text or images upon request. The use of this technology creates a risk of harm to others. Hence the designers of LLM's should be liable if they acted negligently in designing and training the model. In other words, the proper analogy is not to a negligent or reckless journalist or author but to a defectively designed product.

Several different parties may be in the chain of production of an LLM model.[23] One company might produce a foundation model and pre-train it; a second company might fine-tune the model; a third might offer the model as a service to end users. We might analogize their respective liabilities to the different parties who collectively produce a finished product for use by consumers. The duty that each party owes to a defamed party depends on the role that it plays in producing the LLM used by the prompter.

In particular, designers of generative AI systems should have a duty to implement safeguards that reasonably reduce the risk of producing defamatory content. This duty includes a duty of reasonable care in choosing materials for pre-training and fine tuning. It also includes a duty to design and incorporate algorithms that can detect and filter out potentially harmful material, a duty to conduct thorough testing to identify and mitigate risks, and a duty to continually update systems in response to new problems and threats. Traditional product-liability duties to warn should also apply to alert users when a model has an elevated risk of being defamatory – either because it has a heightened risk of being untrue or because it has a heightened risk of being harmful if untrue. Of course, as in standard product liability cases, mere warning does not excuse failure to exercise reasonable care in design. Finally, designers of AI systems are responsible for foreseeable misuses of the systems, because a reasonable designer would certainly be aware that others will use the technology to defame others. ChatGPT, for example, is already programmed not to respond to prompts that are "intended to defame or harm someone's reputation."[24]

Those who use generative AI products to generate prompts can be liable for defamation if they publish the defamatory content that results. The prompter can either be another AI agent or a human being. If the prompter is an AI agent, we should apply a products liability approach similar to that described above.

Suppose, however, that the prompter is a human being who publishes the results to third parties. Human beings, unlike LLMs, do have intentions. They make decisions about what prompts to use and about whether to publish the results of a given prompt. Therefore, the ordinary rules of defamation law should apply, but with the following qualifications.

---

[23] See Katherine Lee, A. Feder Cooper & James Grimmelmann, *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, Journal of the Copyright Society of the USA 31 (Forthcoming 2024) (describing the multiple parties involved in the production and use of LLMs).

[24] David Gewirtz, *6 things ChatGPT can't do (and another 20 things it refuses to do),* ZDNET https://www.zdnet.com/article/6-things-chatgpt-cant-do-and-another-20-it-refuses-to-do/ .

First, suppose that a prompter designs a prompt for the purpose of producing defamatory material about a public figure whether or not it is true. The law should regard this as actual malice.

Second, a reasonable person who prompts an LLM should be aware that an LLM asked for defamatory material may hallucinate in response. Therefore end-users, particularly those who deploy these systems for purposes that carry a high risk of generating defamatory content, have a responsibility to exercise reasonable care both in designing their prompts and in deciding whether to disseminate the results to the public. At some point in the future certain LLMs may be so well designed that their hallucinations are rare and end users who offer prompts to these LLMs can reasonably rely on the results. But under current conditions this reliance is unwarranted. End users therefore have a responsibility to take reasonable steps to verify the accuracy of the content produced by AI, and to refrain from disseminating material that they know, or reasonably should know, is false and defamatory.

A legal framework for defamation law must be adaptable to the rapid evolution of technology without chilling useful innovations. A negligence-based approach offers the flexibility needed to respond to emerging challenges, while holding designers and users liable for failure to exercise reasonable care.

B. Copyright Infringement

Concerns about copyright infringement have led to multiple law suits against OpenAI.[25] These suits raise a number of difficult issues, including whether the material produced by large language models is sufficiently transformative to qualify as fair use.[26] Although intent to infringe is not necessary for copyright liability,[27] proof of intent is relevant to damages. "Willful" infringements are liable for statutory damages of up to $150,000 per work.[28] A defendant who deliberately induces or encourages infringing conduct may also be held liable as a contributory infringer.[29]

---

[25] Ian Kreitzberg, *Here are all the copyright lawsuits against ChatGPT-maker OpenAI*, The Street https://www.thestreet.com/technology/copyright-lawsuits-against-openai-microsoft-chatgpt#:~:text=The%20Authors%20Guild%20of%20America,-The%20Authors%20Guild&text=The%20suit%20claims%20that%20the,copyright%20law%20was%20done%20knowingly.

[26] Lee, Cooper & Grimmelmann, *supra* note 23 (exhaustively discussing various theories of liability).

[27] Buck v. Jewell-Lasalle Realty Co., 283 U.S. 191, 198 (1931) ("[i]ntention to infringe is not essential under the Act"). See also ABKCO Music, Inc. v. Harrisongs Music, Ltd., 722 F.2d 988, 998 n.11 (2d Cir. 1983) ("[I]f copying did in fact occur; [sic] it cannot be defended on the ground that it was done unconsciously and without intent to appropriate plaintiff's work."); R. Anthony Reese, Innocent Infringement in U.S. Copyright Law: A History, 30 Columb. J. Law & Arts 2 (2007).

[28] 17 U.S.C §504 (c)(2) (Revised August 2023).

[29] MGM Studios, Inc. v. Grokster, Ltd., 545 U.S. 913, 930 (2005) ("One infringes contributorily by intentionally inducing or encouraging direct infringement ... and infringes vicariously by profiting from direct infringement while declining to exercise a right to stop or limit it.")

Because LLMs lack intentions, one cannot argue that they willfully infringed. Nor do they have the requisite intentions to induce or encourage copyright infringement. In assigning copyright liability, we believe that the law should focus on whether particular humans acted reasonably with regard to the design and use of technology.[30]

Licensing is one possible solution. LLM companies might secure permission from copyright holders to use their works in training the models and producing new works. Some companies are already securing licenses "either through a bilateral negotiation or by means of an open-source license offered to the world by the dataset compiler."[31] AI companies might try to form a collective rights organization, analogous to the collective performance rights organizations (ASCAP, BMI and SESAC) to facilitate blanket licenses of copyrighted material. Government might help facilitate such a licensing organization.

Licensing systems may not solve all of the problems, however. The success of existing performance rights organizations depends on relative industry concentration. It might be possible to negotiate with the largest commercial owners of copyrights, but training data may include the works of countless smaller producers and individuals, all of whom can make copyright claims. Training data may also sweep up a plethora of orphan works and content produced by authors around the world who will be difficult to locate and compensate. Large copyright-owning companies who are willing to license performance rights for existing works might be unwilling to license the creation of an endless supply of new derivative works that complete with the originals for limited human attention. Conversely, some of the largest AI companies train on end-user data that they collect—and already have licenses for—and they may be unwilling to offer licenses to competitors.

The alternative path is to show that AI companies' use of copyrighted materials is transformative fair use. The doctrinal issues are quite complicated and beyond the scope of this essay.[32] But the deeper problem is that legal doctrine generally thinks of fair use in terms of individual determinations that compare a single derivative work used in a particular way with an asserted original.[33]

This model of individualized determinations of copying or substantial similarity does not make much sense in the context of programs that can combine and recombine any number of elements in their training data and generate an endless supply of new works.[34]

We think that AI will eventually require the law to rethink the premises of fair use doctrine. Put in terms of the basic theme of this essay, AI programs are risky agents; in this case, they create perpetual (and pervasive) risks of copyright infringement at scale. Such infringement is

[30] The supply chain has several subcomponents that can be performed by multiple entities. Lee, Cooper & Grimmelmann, *supra* note 23, at 31. But here we will imagine a unitary actor that develops, trains, and fine-tunes a model.
[31] *Id.* at 38.
[32] See *id.* at 105-114 (exhaustively analyzing the issues).
[33] *Andy Warhol Foundation for Visual Arts, Inc. v. Goldsmith*, 598 U.S. ___ (2023).
[34] The move from individualized to aggregate assessment of reasonableness is also a theme of Omri Ben-Shahar's contribution to this conference. Omri Ben-Shar, Safety Score Liability: A Vision of Tort Law in Era of Artificial Intelligence (working paper, 2024).

problematic to the extent that it undermines incentives for the production of new creative work by human beings. To respond to this problem, the law should require, as a condition of a fair use defense, that AI companies take a series of reasonable steps that reduce the risk of copyright infringement even if they cannot completely eliminate it. The rise of AI may also require that governments find new ways besides copyright law to incentivize cultural production by human beings, but that is a topic for another essay.

A fair use defense tied to these requirements is akin to a safe harbor rule. Instead of litigating in each case whether a particular output of a particular AI prompt violated copyright, as we do today, this approach asks whether the AI company has put sufficient efforts into risk reduction. If it has, its practices constitute fair use.[35]

Here are a few examples of what a safe harbor might require. First, the law might require companies to implement filters that block LLMs from outputting results that copy large chunks of training materials or are substantially similar to them. Some AI programs already implement such filters. For example, GitHub Copilot includes an option allowing users to block code suggestions that match publicly available code.[36]

The problem of matching new digital content to copyrighted works is not a new one. YouTube can automatically detect whether uploaded videos match copyrighted material in their database. Indeed, AI developers necessarily have access to the works on which the program was trained to facilitate such comparisons. AI programs will present a host of new problems of matching text, audio, and graphics. Nevertheless, a safe harbor rule should be to give companies incentives to develop the filtering technology and implement it. The goal, once again, is risk regulation—not to produce filters that work perfectly in every case, but rather technologies that reduce the risks of copying and generating unauthorized derivative works to a reasonable extent.

Filters will also be necessary to deal with prompts that induce copyright violations. Prompts that violate copyright are increasingly likely now that systems like *Claude* allow users to post up to 75,000-word documents.[37] LLM design might automatically check and disable prompts that included unlicensed works contained in the dataset. Generative AI systems already implement a number of prompt filters for unsafe or offensive content.[38] A safe harbor model might require AI providers to install reasonable input and output filters. Finally, it might create an obligation to

---

[35] The Financial Artificial Intelligence Risk Reduction Act, introduced in December 2023 by Senators John Kennedy and Mark Warner, contains an analogous provision that eliminates the scienter requirement and replaces it with strict liability for security law violations unless **"**such person took reasonable steps to prevent such acts, practices, conduct and outcome." S. 3554, 118 Cong. § 42(a) (2023).

[36] *Configuring GitHub Copilot in your environment*, GitHub (Aug. 17, 2023), https://docs. github.com/en/copilot/configuring-github-copilot/configuring-github-copilot-inyour-environment. https://news.ycombinator.com/item?id=33226515 (for related discussion on the Hacker News forum).

[37] Maria Diaz, *4 things Claude Ai can do but ChatGPT can't*, ZDNET, https://www.zdnet.com/article/4-things-claude-ai-can-do-that-chatgpt-cant/

[38] Lee, Cooper & Grimmelman, *supra* note 20 at 48.

update filters and remove content in response to specific requests from copyright owners akin to the takedown requirements of the Digital Millennium Copyright Act.[39]

V. Conclusion

The spread of AI technology will likely require changes in many different areas of the law. In this essay we've argued for viewing AI technology not in terms of its independent agency but in terms of the people and companies that design, deploy, offer and use the technology. To properly regulate AI, we need to keep our focus on the human beings behind it.

[39] Digital Millennium Copyright Act ("DMCA") 17 U.S.C. § 512(c) (3)