Holding AI Accountable: Addressing AI-Related Harms Through Existing Tort Doctrines

Anat Lior*

* * *

I. Introduction

This paper examines the distinct features of artificial intelligence (AI) and reaches a broader conclusion as to the availability and applicability of first-order tort rules. It evaluates the accuracy of the argument that AI is similar in essence to other emerging technologies that have entered our lives since the First Industrial Revolution and, therefore, does not require special legal treatment. The paper will explore whether our current tort doctrines can serve us well even when addressing AI liability.

Traditional tort law doctrines—such as foreseeability and proximate cause, market share liability, and respondeat superior—have yet to be displaced due to AI. Parallel frameworks, such as liability insurance as a risk-hedging instrument, have equally sustained the AI storm. The paper shows why, despite the AI revolution, and notwithstanding the "black box" challenges it generates (which complicate the ability to draw a legal nexus between a human tortfeasor and a given harm), traditional tort doctrines are still extremely relevant, apt, and applicable today. These doctrines are inherently flexible, which is exactly why the tort system has always been able to tackle new challenges resulting from human innovation and emerging technologies.

II. AI Challenges the Tort System

The primary challenge which stands at the center of the AI liability debates is the "black box" issue. This refers to the fact that an AI-based algorithm's decision-making progress is opaque and unknown to the user or programmer of the algorithm. This hampers a core

[*] Anat Lior is an assistant professor at Drexel University's Thomas R. Kline School of Law, an AI Schmidt affiliated Scholar with the Jackson School at Yale, and an affiliated fellow at the Yale Information Society Project. Her research focuses on Artificial Intelligence and its interaction with tort law, insurance law, and antitrust law. She commonly confronts issues such as AI regulation and policy, AI liability, and insurance as applied to emerging technologies. She would like to thank Asaf Lubin, Jessa Feiler, and the participants of "How AI Will Change the Law" symposium for their helpful comments.

notion standing at the heart of tort law—foreseeability and predictability. These are required to prove the causes of actions rooted in product liability, negligence, and strict liability. Companies developing AI often claim they cannot know how the algorithm reached a harmful decision or recommendation, as it does not reveal the proxies it uses in its decision-making process. Some claim that this might prevent AI from fitting into our existing liability doctrines because these legal doctrines were designed with human conduct in mind; thus, they might not function when applied to AI. Furthermore, when discussing negligence and design defects liability regimes, scholars have argued that those cannot apply in the context of AI liability as the notion of "fault" becomes murky when AI makes its "own" harmful decisions.

Despite this, tort law has been well known throughout history for adopting and adjusting itself optimally to novel technological risks entering the commercial market. We have seen this since the days of the First Industrial Revolution, which eventually led to a slow shift to a negligence regime and the creation of a workers' compensation system. We also saw this happen when the automobile was invented—the tort system was adapted to handle the carnage that is still ongoing on the roads. More recently, we have seen this in cyberspace, where the tort system was indeed adjusted to accommodate the risks and harms produced by cyberspace but was not replaced or reinvented.

Embedded in tort law is the ability of decision-makers, both judges and juries, to balance the competing interests entrenched in the assimilation of any new technology into our commercial market—consumer safety, judicial efficiency, and the support of a new industry. There is no reason to believe, nor is it supported by evidence, that the tort system will not be able to do the same regarding the AI industry. Even if we adjust certain aspects of current legal tort doctrines, such as modifying the standard of care to incorporate AI development and usage, this still does not render our legal infrastructure inept.

The next part presents different first-order tort rules that could and should apply when AI liability arises. The ongoing claim that AI should completely alter the current tort system is not supported by evidence in practice. Though some rules might need adjustment, they were all created with flexibility in mind and, as such, are capable of tackling AI-causing harms. This is especially true given the legal fallacy that AI entities should be viewed and treated as independent agents without principals, as Section IV will elaborate on.

III. First-Order Tort Rules and Their AI Applicability

This part delves into three first-order tort doctrines—foreseeability and proximate cause, market share liability, and respondeat superior—as well as the product of liability insurance. Regarding liability insurance, though not a traditional first-order tort rule, it is an essential underexplored aspect of the tort system. It holds great value in the context of AI, as will be elaborated below, to nudge policyholders to behave safely while engaging with AI. This is due to the current legislation void left by regulators who are still struggling with the appropriate AI-policy approach.

With regard to the three traditional first-order tort rules, foreseeability and proximate cause have risen as significant challenges in the context of AI liability, whereas market share liability and respondeat superior represent legal doctrines that seem extremely apt in the AI world. AI is challenging many other first-order tort doctrines; these three were chosen to provide a glimpse into this debate. Foreseeability and proximate cause are the most highly discussed notions that are being pushed against in the context of AI. Market share liability and respondeat superior, though they could be useful in the AI context, have been rejected or simply underexplored so far.

A. Foreseeability and Proximate Cause

The doctrine of proximate causation essentially examines whether a specific harm was a foreseeable consequence of a defendant's conduct. It appears that legal scholars working on AI liability have largely sidestepped causation issues in their writing. This is because deciding whether a given accident involves AI or was caused by AI is extremely challenging. However, sidestepping or rejecting this element does not seem appropriate, given its significance in maintaining fairness and efficiency within the tort system.

AI risks can be roughly divided into three main categories: First, alignment failures, where the AI has a different set of goals than those provided by humans. Second, capabilities failures, where the AI malfunctions. Third, misuse, where the AI functions as planned but with the human creators' malicious intentions. The misuse category seems to fall squarely under intentional torts or negligence, which should not be difficult to prove. The capabilities category poses no difficulty to the foreseeability question because usually, when the AI malfunctions, the resulting inflicted damages should have been expected and are considered foreseeable risks associated with the malfunction.

Alignment failures present a more challenging issue in the notion of foreseeability. This is because the proximate cause question will depend on the level of generalization under which the concept of foreseeability is examined. The generalization problem is again not unique to the AI context, and the level of generalization the fact-finder decides upon will determine the proximate causation issue. Alignment failures are indeed foreseeable, and more experts warn us about them each day. However, it is extremely hard to foresee the specific alignment failure that could emerge. Though it is unclear how courts will interpret foreseeability in misalignment cases, it seems only fair to apply a high level of generality, as AI companies know this problem is inherent to this emergent technology and are still choosing to disseminate it commercially. If indeed cases will turn on the level of generality applied, the proximate cause is still very relevant and applicable. The only way to achieve the function of optimal deterrence is to apply a high level of generalization while examining the foreseeability of a given harm. Otherwise, AI companies and users will claim that the alignment problem prevents the assignment of liability altogether.

The existing doctrine of proximate cause gives our common law system the discretion to apply the concept of foreseeability at a high level of generality when a misalignment issue lurks in the background. The Restatement (Third) of Torts (Liability for Physical and Emotional Harm) explicitly leaves this issue to the factfinders' "judgment and common sense." Specialists in different fields, such as healthcare, can use common sense and logic to "crack" the stern exterior of AI's "black-box" problem to at least offer foreseeable harm as AI is being used for specific purposes. This could provide factfinders with a better understanding and background of the possible harms that can happen once AI is deployed.

Professor Nick Bostrom and Luke Muehlhauser of Open Philanthropy have even famously claimed that it is somehow predictable that during our quest to maximize the creation of paperclips by an AI entity, an apocalypse might ensue as the AI will divert all possible resources to complete this task, thus mining the earth itself. Though still science fiction at the moment, this example can be used to show that it is certainly foreseeable that AI will cause catastrophic risks and that those creating and using AI should consider this when deciding to develop and use it.

The jury does not have to decide that the specific way the damage happened was foreseeable if the harm itself was a result of the general type of risk that a reasonable person should have taken steps to mitigate. Given what we know about the alignment problem, it seems sufficient to find that a general sort of risk exists, even if we cannot precisely predict how this risk will come to be. The inherent flexibility of the proximate cause doctrine will successfully apply to cases of AI-inflicted harm. As a result, the doctrine of proximate cause should not be changed or adjusted because it was meant to act as a "safety valve" regardless of the domain it is being applied in, even if that domain is AI. This connects to the notion of using inference-based analysis to determine liability in AI-related injuries, even if we don't have direct evidence of fault.

B. *Market Share Liability*

Moving from proximate cause, this Section delves into the factual aspect of the causation element (the but-for test). It acknowledges that it is sometimes difficult to pinpoint the entity that fulfills the causation-in-fact, but even then, tort law is familiar with these challenges. This Section directly connects to my comment above about legal scholars sidestepping the causation problem altogether.

The doctrine of market share liability can be valuable in dealing with damages caused by AI when it is unclear who the liable entity within the AI industry is. This is especially true given the highly condensed structure of this industry, which is dominated by a few, giant, primarily U.S.-based tech companies. Focusing on the substantial share of the market held by these companies, there could be future scenarios where this doctrine will be required to establish causation in-fact. Courts have indeed been reluctant to apply this doctrine outside of the diethystilbestrol (DES) context, focusing on whether the product is fungible (i.e., "the manufacturers acted in a parallel manner to produce an identical, generically marketed product") and on the manifestation of the injury being far removed from the time the product was used.

Given the scale of AI, the similarity of algorithms based on machine learning and large language models deployed by different AI companies, the way the industry is structured and rapidly developing, and AI's ability to cause damages that will only manifest in the future, the application of this doctrine seems appropriate. Utilizing it can also incentivize AI companies to document and track the AI-based products they disseminate to the public. Market share liability could help prevent future cases where consumers cannot sue for harm because they are unsure which company created the AI that harmed them.

Scholars have proposed applying market share liability to hold tech companies accountable if they are found to create cyber nuisance towards their users, as well as in the autonomous vehicles context when “the entire industry of a particular type of autonomous vehicle could be assessed liability for all the damages arising from all the unavoidable accidents involving that particular type of vehicle.” The moral basis for these proposals, which is the moral justification for this doctrine in the first place, is that within an industry where companies learn from each other’s product experiences and operate under the same safety standards, a network effect emerges, making all entities partly responsible if a flawed product enters the market.

The applicability of the market share doctrine seems especially apt in the context of long-tail liability cases where AI technology is used today, but the damage it inflicts will only be apparent after years. These scenarios are similar to the circumstances that led the *Sindell* court to establish this vital doctrine. This could apply, for example, in cases regarding AI-assisted CRISPR, gene editing, and even certain facial recognition harms that can go undetected for years.

C. Respondeat Superior

I have written about the applicability of the respondeat superior doctrine in the AI context at length elsewhere. This doctrine is based on the notion that if a principal entrusts subordinates (*e.g.*, an AI entity) to carry out an inherently risky activity, then principles of fairness necessitate that the principal bear responsibility for that conduct if it results in harm. In practice, one can claim the principal is in a better position to bear the costs of the damage or acquire insurance against potential liability claims than the agent. This can also motivate the principal to choose its agents more carefully and invest more time and effort in their selection, training, and subsequent monitoring. This doctrine is designed, in part, to ensure tort victims will not be undercompensated in the case of an insolvent agent. This rationale is fundamental when we discuss AI entities because they are inherently insolvent. AI entities are neither humans nor corporations and thus have no pockets from which to pay.

To establish this doctrine, there should be an appropriate relationship between the superior and its subordinate, and an appropriate connection between that relationship and the conduct of the subordinate that led to damage. To decide whether respondeat superior should apply in a specific case, courts commonly examine whether the agent was acting “in the course of the employment” when the damage occurred. This is meant to distinguish between acts carried out by the agent for which the principal will not be held liable (*e.g.*, frolics) and those for which it will.

In the AI context, this distinction does not exist, as there are no acts that can be carried out by the AI agents that will exceed the liability scope of the principal. This is because AI agents have single or multiple purposes pre-coded in them by their human creators, even if they act in misalignment to the assigned task(s). AI agents are functional entities created to serve a single or a set of functions. The entire purpose of their existence is to carry out the tasks coded into them. They have no frolics to engage in; they have no detours to take.

The main challenge here is determining the AI agent's principals. The principals should be identified as those with the highest capability to affect the actions of an AI entity through monitoring, supervision, and guidance. The AI entity is in a constant state of "being an agent" for the benefit of others, mainly of its principals. The question is who the most significant "pressure point" is and will be willing and able to consider the costs its AI agent may inflict. Based on these costs, humans or corporations—principals—are in the best position to decide whether to better equip and train the AI agent or to pay the price in the form of monetary sanctions or an insurance premium.

The identity of the principals will change per instance and will heavily depend on the circumstances of an accident. This includes considering different factors, such as the level of involvement, supervision, monitoring, and ability to direct the actions of the AI agent given the inflicted damages. At the early stages of an AI agent's development, this level of control will be frequently attributed to the designer, programmer, trainer, or manufacturer of the AI agent rather than its operator or owner. The more the usage of these AI agents becomes pervasive, however, the more likely the operator's or owner's level of control and monitoring will result in identifying them as the appropriate principals.

There will be cases where an AI agent inflicts damage while under the control and guidance of more than one principal. Multiple principals can be viewed as joint principals who are in co-control over the AI agent's activities. When more than one entity can be identified as the AI's principal, all relevant principals should be held liable for the damage that occurred jointly and severally.

D. Liability Insurance

Though not a first-order tort rule, it is important to note the instrument of liability insurance and its potential role in the context of AI liability. Liability insurance has had an important, though underappreciated, impact on the development of tort law. The same trajectory of liability insurance influencing the tort system is certainly possible in the context of AI liability, as more and more insurance and reinsurance companies are offering policies to cover risks associated with AI (*e.g.*, Munich Re and Vouch).

In my previous work, I argued that the discussion about AI liability has primarily focused on the appropriate liability regime rather than considering the policy implications of liability insurance that will flow from such regimes. While we are still in this adjustment period where more is unknown than known about AI capabilities, insurance can help avoid legal issues of blame-placing and provide much-needed compensation to those harmed by the deployment of AI. Insurers can incentivize the behavior of their policyholders (via their different products) to act cautiously once AI is involved. Liability insurance can enhance the integration of AI into daily commercial routines while mitigating the harms that may arise from this process.

As the risks associated with AI become more familiar, insurance companies will have the necessary datasets to calculate accurate premiums and provide valuable loss prevention services to their policyholders. This will also enable this technology to be implemented more safely until the courts can decide the best approach toward AI-liability scenarios.

Liability insurance is not a stand-alone solution and has many negative implications once involved in a specific industry. These include moral hazards, adverse selection, regulatory capture, negative externalities, and much more. Nonetheless, it is an integral part of the overall tort approach policymakers should consider. It should be considered across different AI applications to ensure innovation can be supported while potential victims are compensated for their losses. This should be done while tackling the dark sides of the insurance instrument and industry.

IV. AI as An(y) Emergent Technology

In *The Wizard of Oz*, when Dorthy and her merry bunch discover that the Wizard is just a man and not the series of flashes, loud noises, and flames that were projected to them, the Wizard's first response is, "[p]ay no attention to that man behind the curtain." In many ways, this reaction conveys the essence of how big tech companies view their AI outputs when acknowledging the possibility of harm—*pay no attention to us*.

A recent attempt by Air Canada to shrug off a commitment to a discount made by their AI chatbot to a customer demonstrates this approach. Air Canada argued that its chatbot is a separate legal entity that is responsible for its *own* actions. In *Moffatt v. Air Canada,* a Canadian tribunal rejected this argument and obligated Air Canada to provide the discount as promised by the chatbot, essentially lifting the curtain between an AI entity and the legal entity standing behind it. In the context of price fixing, the FTC recently published a joint legal brief with the DOJ stating that "your algorithm can't do anything that would be illegal if done by a real person." Assuming liability is established, this logic also applies in tort law.

New technologies have always put the tort system under constraint. Since the days of the First Industrial Revolution, through the innovation of the hot air balloon, the automobile, the airplane, and the Internet, it seems that every new technology has triggered alarm bells. These alarmist warnings echoed the same claim: the tort system is on the verge of collapsing and should thus be changed or completely altered to address the new technological challenge. However, the tort system has been able to handle these technological developments without significantly reinventing itself.

The solutions for confronting AI liability should be generated from within the system without the need to rewrite it altogether. The tort system usually reacts with suspicion to new technologies because it cannot achieve optimal deterrence at first, a vital function of this system, as it lacks information about the harms new technologies can inflict. As a result, it tends to impose a rigorous liability regime at first, in the form of strict liability. As the social and economic benefits of the new technology become apparent, as well as its associated risks, the tort system tends to ease its grasp and shift to a more flexible and accommodating regime in the form of negligence or safe harbors. This natural cycle seems apt in the context of AI given the black-box issue and the "known unknown" risks ("contingencies that we know exist, but to which neither a probability nor a magnitude can be actuarially assigned") associated with this cross-disciplinary technology.

In this sense, AI is similar to previous emergent technologies that have changed our perception of what is a foreseeable risk and what risks we are willing to accept while living in a modern world. The safer new technologies become—because they have been around long enough to understand how and when they inflict harm and what can be done to mitigate these harms—the more comfortable society becomes with these new technologies. In some situations, these technologies become so entrenched in our lives that we cannot imagine ourselves living without them. AI should go through this cycle like its predecessors, and if and when it becomes safe enough, a negligence regime should be more apt once we can better identify what damages are foreseeable from both a factual and normative standpoint.

Conclusion

Not all existing tort doctrines can readily apply to the AI context. A good example of that is using a traditional negligence standard to harm caused by an AI algorithm. Applying the reasonable-person standard via the Learned Hand formula to establish a breach of duty (i.e., when B < P * L) when AI leads to damage is extremely difficult—not impossible, but difficult. Currently, there is no explicit agreement about the precautionary measures (B) one can take in the AI context, as well as the likelihood (P) and severity (L) of potential AI-inflicted harms. Thus, it is hard for factfinders, both judges and juries, to determine what should have been the appropriate level of caution required by a reasonable care duty. This does not mean we should change the elements and features of a negligence cause of action; it simply means that, for now, it is not the appropriate means to measure liability when AI causes damages.

The technology of AI is aimed at conducting precisely what humans are doing—driving, performing surgery, writing legal memos, and trading stocks—in a more efficient and hopefully safer manner. We already have legal doctrines to govern these types of behaviors and the damages that can result from these activities. An attempt to reinvent the tort wheel once AI is widespread could lead to confusion and uncertainty. This argument is amplified given the apparent lack of expertise of the legal representatives in charge of crafting legislation to govern AI. Courts may decide to develop new tort doctrines as AI capabilities become clearer and more advanced, such as AI personhood, but this does not render current legal doctrines moot or irrelevant.

* * *

Anat Lior is an assistant professor at Drexel University's Thomas R. Kline School of Law.