Regulating Artificial Intelligence: Discouraging misuse without hampering innovation

Robert Mahari and Alex Pentland, MIT
8/17/2021

This abridged paper provides a summary of our forthcoming article on strict liability for AI. We welcome your feedback. Please treat confidentially.

I. Introduction

Civil law provides two fundamental approaches to regulating harmful behavior. The first is ex-ante intervention, that seeks to *prevent* the behavior that gives rise to harm. The second is ex-post liability, that retroactively punishes harmful conduct. The past decade has witnessed a shift towards a preventative approach to regulating information technologies in an effort to protect consumers from unsafe products.[1] However, application of this well-intentioned philosophy to artificial intelligence (AI), does little to prevent harm while threatening to hamper innovation. Given the nebulous definition of AI in the first place, and the rapid pace at which the field develops, ex-ante regulation confronts lawmakers with the impossible task of predicting the trajectory of the AI industry. This challenge is compounded by the distributed and international nature of the AI development community, characteristics that make ex-ante regulation impractical to enforce. In our view, an ex-post liability regime that punishes AI misuse is likely a more realistic and efficient legal framework. We advocate for an approach that shifts the costs resulting from AI harms to manufacturers and thereby incentivizes the latter to deploy safe products and to self-regulate. These objectives are achieved by the strict liability regime that has been applied to products in the United States and elsewhere since the 1960s.[2]

In this paper, we explain why an ex-ante regime is a poor fit for AI and explore how a strict liability regime might be applied to AI, allowing citizens to and government to ensure citizen rights and mitigate harms. To implement this shift requires that manufacturers and service providers keep a detailed record of all decisions made by an AI (analogous to requiring companies to keep financial records), creation of regulatory bodies (analogous to financial regulators) that are able to regularly audit these decision records in order to ensure citizen rights, and creation of trusts (analogous to credit unions) that are able to hold personal data for citizens, advocate for their rights, and aid them in litigation to mitigate harms.

II. Why ex-ante regulation is the wrong general approach for AI regulation

[1] Zittrain, Jonathan, Three Eras of Digital Governance (September 23, 2019). Available at SSRN: https://ssrn.com/abstract=3458435
[2] *See* Greenman v. Yuba Power Products, Inc. 377 P.2d 897 (Cal. 1963)

As jurisdictions around the world grapple with the challenge of formulating AI regulation, legislators seem drawn to the idea of creating a legal framework that ensures only "safe and trustworthy" AI enters the market.[3] This regulatory approach betrays a fundamental misunderstanding about what AI is and who builds AI. In this section we argue that ex-ante regulation is, generally, a poor approach to AI regulation for three reasons. First, AI is a broad and evolving field which neither lends itself to one-size regulation nor permits accurate predictions about future developments. Second, AI development features low barriers to entry which promote a distributed and international AI development community, making the enforcement of ex-ante regulation unrealistic. Finally, ex-ante regulations not only fail to protect consumers, but also needlessly hamper innovation.

2.1 Regulating the full spectrum of AI ex-ante is inherently unrealistic.

As a result of its multifaceted and rapidly changing nature, AI does not lend itself to one-size ex-ante regulation. Artificial intelligence is a broad term that has been poorly understood and hence ill-defined by regulators. For example, proposed EU regulation on AI defines AI very broadly to include techniques ranging from deep machine learning approaches to basic statistics and Bayesian estimation.[4] Fundamentally, AI systems are a method of capturing patterns in data and utilizing the resulting insights, and the term "AI" captures systems ranging from relatively innocuous tax expert systems to facial recognition surveillance systems with far reaching consequences. Similarly, to other broad categories of products – like consumer goods – AI systems come in many shapes and sizes, but the types of harm that they may cause, and which we seek to prevent, are relatively easy to enumerate (see section 3.1.3). Ex-ante regulation is misguided because it is focuses on the complexity of AI systems and how they are developed, rather than on the tractable set of possible harms these systems are capable of.

2.2 The distributed and international nature of AI makes ex-ante regulation difficult to enforce

As a result of the relatively low barriers to entry, AI development is not centralized in major entities but rather a distributed international pursuit which makes enforcing ex-ante regulation unrealistic in practice. The basic ingredients for AI are a model, training data, and compute resources needed to train the model. AI models may simply be a mathematical procedure which can be expressed as code. Such models are described in academic papers and have been made available through popular free packages like PyTorch[5] and Keras[6]. Increasingly, practitioners rely on pretrained AI models that have been pretrained on large amounts of data; these pretrained models are generally also freely available. Many of these models can be trained on regular computers, but more powerful computing power is made available for free by Google Colab[7] or

[3] *See* REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS (2021). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
[4] *See supra* note 2
[5] https://pytorch.org/
[6] https://keras.io/
[7] https://colab.research.google.com/notebooks/gpu.ipynb

at low costs for other provides. Obtaining training data is likely the biggest bottleneck in the AI development process as the type and quantity of data needed is very task specific. Nonetheless, many datasets can be obtained freely online or acquired relatively cheaply through the growing number of training data vendors.

The availability of these resources, together with the vast amount of accessible instruction on AI development found in online blogs, makes the barriers to entry for AI low. As a result, the AI community contains many small players that operate globally. These characteristics make the enforcement of ex-ante regulation challenging and costly. While proposed ex-ante regulation purports to ensure only safe AI reach the marketplace, in practice this regulatory approach would likely be unable to monitor what AI systems are deployed at all and would inevitably be rendered reactionary.

2.3. Enforcing ex-ante regulation hampers innovation

Both EU and US government guidelines stress the need protect consumers without hampering AI innovation.[8] As outlined above, ex-ante regulation does little to protect consumers from harm, but it also imposes significant burdens on AI development. For example, proposed EU regulation on AI requires that, for high-risk AI systems, "[t]raining, validation and testing data sets shall be relevant, representative, free of errors and complete". Requirements of this nature are inherently unrealistic, as virtually no dataset is likely to be complete and some degree of error is always present. Imposing ex-ante regulations that misunderstand how AI is developed not only fails to protect consumers but also deters desirable innovation.

III. Designing a framework for AI liability

For the reasons outlined in the prior section, we believe that attempting to use regulation to predict and prevent AI harms ex-ante is ineffective. Instead, we advocate for a legal framework that introduces incentives for AI creators to release safe products. To this end, we advocate for a solution grounded in tort law that draws on well-established strict product liability principles. Applying strict liability to AI would mean that AI manufacturers are liable for harms caused by defective or falsely advertised AI systems. This section outlines some key considerations for applying strict liability to AI.

3.1 AI Liability Framework

Tort law provides a legal basis to shift losses associated with unsafe conduct from the injured party to the person causing the harm, hereby discouraging unsafe conduct.[9] In most cases, to win in a tort suit, the injured party must demonstrate that the alleged tortfeasor acted negligently.

[8] *See supra* note 2; MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES (2020). Available at https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf?utm_source=morning_brew

[9] George L. Priest, Satisfying the Multiple Goals of Tort Law, 22 Val. U. L. Rev. 643, 648 (1988). Available at: https://scholar.valpo.edu/vulr/vol22/iss3/5

However, in the case of product liability, U.S. law generally applies the principle of strict liability meaning that the injured party need only show that her injuries arose from a product that was defective or falsely advertised. We believe that applying the principles of strict liability to harms caused by AI creates effective incentives for AI manufacturers to release safe products.

Under a strict product liability regime, manufactures are liable for harms to persons or property caused by defects in the products they sell, where a defect may be a manufacturing defect, a defect in design or a failure to properly warn consumers.[10] The following discussions will begin to explore what constitutes a defective AI system, when an AI system "causes" a harm, and what types of harm should be included in a strict liability framework for AI.

3.1.1 Defective AI

This paper, like most regulation on AI, focuses on "weak" AI systems, that have a narrow and well-defined application.[11] In cases where AI manufacturers clearly state how their AI operates and what level of accuracy may be expected, a defective system will often be obvious. For example, an AI system used by a vegetable wholesaler to identify spoilt produce might have an acceptable error rate and if the AI performs significantly below this rate, then this may constitute a defect. Generally, if a manufacturer fails to warn consumers about a significant limitation, she may be liable for damages resulting from the improper but foreseeable use of the AI. However, while it is desirable for manufactures to clearly state their systems' limitations, disclosure mustn't become a carte blanche that allows AI manufactures to escape all consequences. To this end, limitations that could cause harms during the reasonably foreseeable use of an AI system should be regarded as defects, regardless of whether they have been disclosed. For example, an AI resume screening system that systematically rejects older applicants should likely be deemed defective regardless of whether the manufacturer disclosed this shortcoming. On the other hand, if the AI system presents a problem that the manufacture did not disclose, this should not automatically constitute a defect. In these cases, it may be helpful to compare a given AI system to another system to establish a baseline. For example, an AI resume screener that rejects numerous qualified candidates may be defective if it is advertised as a replacement for human review but nonetheless rejects many candidates that a human reviewer would not. This nuanced distinction raises an ethical question in instances where an AI is to replace a human process which is not performing well to begin with. In our view, unless AI systems are deployed in a limited number of high-risk applications (see section 3.2), a reduction in system error should be viewed as positive and as a sign that the AI in question is not defective.

3.1.2 Causation for AI harms

[10] RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. §1 – 2 (1998): One engaged in the business of selling or otherwise distributing products who sells or distributes a defective product is subject to liability for harm to persons or property caused by the defect… 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…
[11] As opposed to strong AI which can complete a multitude of tasks.

AI systems are often criticized for their lack of explainability and this shortcoming may make it challenging to determine when a defective AI system *caused* a harm. While concerns over AI explainability are sometimes exaggerated – after all, human decision-making is hardly a fully explainable process – a lack of explainability may obscure defects and make causation difficult to pin down. However, it may not be necessary to explain an AI's actions in general to determine whether the AI caused a specific harm. In some cases, it may be enough to compare the AI system to a baseline (a human or another AI) to determine whether the baseline system would likely have given rise to the same harm. In other cases, emerging computational techniques such as LIME could help to explain the reasoning behind specific predictions made by an AI system,[12] and such approaches can help determine the degree to which a model prediction gave rise a certain harm. To enable such ex-post algorithmic auditing, AI manufactures will, at the very least, need to maintain anonymized records of input data and model outputs that can be analyzed during trail. Another dimension of AI causation is the degree to which the AI system can make autonomous decisions. If the AI system's outputs are one factor of a larger decision-making system, then proving causation also requires showing that a different output would have ultimately resulted in a different outcome.

Maintenance of input and output data is key to both harms litigation and vigilance against harm by use of continuous audit. This will likely not be burdensome even for small firms; it is analogous to the burden of keeping accurate financial records. To avoid litigation and improve performance, companies would want to continuously check their performance internally, much like balancing the books each night, but also would be subject to regular outside audit and audit during trial discovery. Regular audit and maintenance of input-output records also permit much more accurate, timely restitution for people who are harmed, and has the additional benefit that over time there should be sufficient diversity of experience to craft carefully targeted ex ante regulation.

3.1.3 AI Damages

Another challenge that must be resolved to effectively apply the strict liability framework to AI is the assessment of damages. Most U.S. courts limit the application of strict product liability to cases where a product causes personal injury or damage to property.[13] In some cases, AI may cause personal injury, for example, proposed EU regulation for AI articulates a worry about AI enabled "subliminal techniques" that induce vulnerable groups to behave in a manner likely to cause "psychological or physical harm". Harm of this nature can be categorized as personal injury and would be captured by a conventional strict product liability framework. Most courts argue that strict product liability should not apply when losses are purely economic, and that

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778

[13] Cathy Bellehumeur, Recovery for Economic Loss Under a Products Liability Theory: From the Beginning Through the Current Trend, 70 Marq. L. Rev. 320 (1987). Available at: http://scholarship.law.marquette.edu/mulr/vol70/iss2/5

contract law should govern in these cases. However, in the case of AI it is likely that the person experiencing the loss caused by a defective AI product does not have a contractual relationship to the AI manufacturer. For example, a job applicant whose application is rejected by a defective AI algorithm incurs a loss but would be unlikely to have a direct contractual relationship with the AI manufacturer. In these cases, it may be possible to assign liability to the firm deploying the AI algorithm, but this firm will often be powerless to modify the AI. For this reason, it may make sense to expand strict liability for AI to also include economic losses when they are caused by an AI system and there was no contractual relationship between the injured party and the manufacturer. Finally, many fears related to AI are not based on personal injury, property damage or even economic losses but instead more intangible damage – like a reduction of privacy. A comprehensive AI liability framework may require the definition of statutory damages to capture these additional harms to ultimately incentive AI manufactures to bear them in mind as they deploy their products.

3.2 Preventing irreparable harms

A small fraction of AI applications will inevitably present the risk of causing significant irreparable harm and in these instances ex post facto relief would be inadequate. In our view irreparable harm is only likely in instances when AI is deployed in a high-risk context, such as law enforcement, social credit scoring or surveillance, and we therefore suggest imposing additional safeguards in these contexts. This section will outline two examples of such safeguards. First, plaintiffs could be given the ability to seek temporary injunctive relief to temporarily block the deployment of high-risk AI that has the potential to cause irreparable harm. The purpose of these temporary injunctions would be to provide a window of time in which a given AI algorithm may be scrutinized closely to determine whether it is likely to result in significant irreparable harm. Second, AI manufacturers may be mandated to meet requirements related to algorithmic transparency and auditability if they seek to deploy high risk AI. Not only would this facilitate the auditing of these systems if a temporary injunction is granted, but it would also likely make it easier for plaintiffs to seek ex post facto relief by reducing the burden associated with proving causation. In this vein, AI systems operating in high-risk domains could be required to adopt principles that seek to ensure fair and transparent algorithms, such as the open algorithms (OPAL) paradigm,[14] which proposes ex-ante algorithmic vetting to ensure that proposed algorithmic systems are privacy-preserving, fair and free from bias.

IV. Conclusion

As AI becomes ever more ubiquitous, regulators seek to design laws that protect consumers while promoting AI innovation. Many regulators are drawn to ex-ante regulation that aims to prevent harms. This paper argues that this ex-ante approach is inferior to an ex-post strict liability regime. Ex-ante regulation is a poor general approach to regulating AI for three reasons.

[14] Hardjono, Thomas, and Alex Pentland. "Open algorithms for identity federation." Future of Information and Communication Conference. Springer, Cham, 2018.

First, AI is a diverse and rapidly changing field that is does not lend itself to one-size regulation or allow accurate predictions about its future trajectory. Second, the AI community is global and composed of many small players, making the enforcement of ex-ante regulation unrealistic. Third, while ex-ante regulation fails to protect consumers, it imposes considerable costs on AI developers and hampers innovation. In lieu of ex-ante regulation we propose a strict liability framework in which AI manufactures are liable for harms caused by defective AI. We discuss several key considerations related to imposing such a regime and explain how it can incentivize AI creators to release safe products.