A Penalty Default Approach to Preemptive Harm Disclosure and Mitigation for AI Systems

A Penalty Default Approach to Preemptive Harm Disclosure and Mitigation for AI Systems

Rui-Jie Yew
rjy@mit.edu
MIT
IDSS, CSAIL
Cambridge, MA, USA

Dylan Hadfield-Menell
dhm@csail.mit.edu
MIT
EECS, CSAIL
Cambridge, MA, USA

ABSTRACT

As AI industry matures, it is important to ensure that the organizations developing these systems have sufficient incentives to identify and mitigate risks and harm. Unfortunately, the profit motive is often misaligned with this goal. Successful work to identify or reduce risk rarely has direct tangible benefits. In this paper, we consider the use of regulatory penalty defaults as a way to counter these perverse incentives. A regulatory penalty default regime consists of two parts: a regulatory penalty default and a mechanism to bargain around the default. The regulatory penalty default induces private actors to research and mitigate potential harms in order to limit liability, making the benefits of risk mitigation tangible. The bargaining mechanism provides incentives for companies to go beyond achieving a prescriptive threshold of compliance in creating a compelling case for escape from the default. With a focus on the policy landscape in the United States, we propose and discuss potential regulatory penalty default regimes for AI systems. For each of our proposals, we also discuss accompanying regulatory pathways for the bargaining process. While regulatory penalty default regimes are not a panacea (we discuss several drawbacks of the proposed methods), they are an important tool to consider in the regulation of AI systems.

CCS CONCEPTS

• **Social and professional topics → Computing / technology policy**.

KEYWORDS

artificial intelligence law, technology policy, computing and society

ACM Reference Format:
Rui-Jie Yew and Dylan Hadfield-Menell. 2022. A Penalty Default Approach to Preemptive Harm Disclosure and Mitigation for AI Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3514094.3534130

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3534130

1 INTRODUCTION

AI practitioners often describe their organization's approach to harm mitigation as "reactive" to external pressures, such as potential negative public relations (PR) and other media exposure [34]. Furthermore, employees who study and mitigate the harms of AI systems often do so as volunteers on top of their other full-time defined roles in the company. In order to garner support for their work, they often need to rely on the possibility of external threats like "what if ProPublica found out?" [34]. From the perspective of an organization's selfish incentives, this makes sense. Foreseeability defines the outer bounds of legal culpability [38]. As a result, any knowledge generated in the often costly effort to identify possible harms is a dubious value proposition. The ability to anticipate harms creates legal responsibility and liability that is only balanced by the *potential* mitigation of bad publicity.

In the current self-regulatory policy regime for AI systems in the United States, PR incentives are central. Thus, the value of an employee who generates information about and/or mitigates harms is hypothetical or, at best, retroactively substantiated by reference to PR disasters that might have been prevented[1]. It is nearly impossible to reify the value an employee contributes through the study and mitigation of harm: the mark of a job well done is that no additional costs are incurred by the potential harm that was prevented.

An alternative to self-regulation is a top-down, "command-and-control," regulatory regime. The drawback of this approach lies in the contextual nature of the development and deployment of AI systems [36, 45]. Command-and-control regulation leaves ample room for unanticipated harms or behavior. It is challenging to translate ethical principles about AI systems into practical guidance for their development and deployment [29]. Despite relatively broad agreement on the importance of abstract principles, such as fairness and accountability [14, 23, 27], it remains difficult to determine how biases in AI systems will manifest across the wide variety of intended applications [37]. Gaps in the actual implementations of these principles are left to be interpreted and filled by practitioners on contextual bases [26]. This gives practitioners a practical and informational advantage over regulators. The organizations that develop AI systems have ample opportunities to subvert the intended goals of regulation. This information asymmetry also means that practitioners are often in the best position to identify and mitigate potential harms. Thus, it is crucial for external policy mechanisms to push organizations to invest appropriate resources in harm identification, disclosure, and mitigation.

[1] Thanks to Chris Jones for this point.

In this paper, we consider the use of *regulatory penalty default regimes* to fill this regulatory need. A regulatory penalty default regime consists of two components: a regulatory penalty default and a bargaining mechanism. A regulatory penalty default is a backdrop regulatory requirement that creates broad penalties for misbehavior (i.e., harms caused by AI systems). The bargaining mechanism is a clearly specified path for private actors to work around the default. For example, this might involve disclosing potential harms and mitigation measures to a regulatory body. This unsavory baseline of prescriptive regulation pushes companies to achieve compliance through commensurate measures to escape the default, not just advocate for the implementation of corporate-friendly policies across the board. On the other hand, the bargaining mechanism moves the burden of proof about the safety of a system from regulators to private actors. It places the onus on companies to prove that they are managing risks well, instead of requiring governments to prove wrongdoing.

Regulatory penalty default regimes create concrete incentives for private employees to look into potential harms. With the bargaining process, private employees can point to how their identification and proposed mitigation of harms allowed the company to escape the penalty default. Crucially this includes harms that may *or may not* be covered by prescriptive regulation. Unlike traditional safe harbors, which typically contain explicitly written conditions and cover a specified range of circumstances, a company's escape from the penalty default depends on its ability to present a strong case for the sufficiency of the mitigation measures in place. Hornstein [13] notes that regulatory penalty defaults can be viewed as a regulatory tool that straddles the exploration-exploitation tradeoff, where it imposes "a known baseline of prescriptive regulation (exploitation) against which the parties would have an incentive to bargain around in developing a less expensive, more innovative means of reaching the agency's goals (exploration)."

Regulatory penalty defaults are not new. They have been applied or proposed in several different contexts. Karkkainen [18] identifies regulatory penalty defaults as information-forcing and action-forcing components of environmental regulation. Kaminski [15] suggests the use of penalty defaults to spur private ordering in multistakeholder privacy governance. Kaminski [16] and Selbst [37] propose penalty defaults as a part of a collaborative governance regime for algorithmic decision-making systems and as part of an algorithmic impact assessment regulatory approach, respectively. In this paper, we build on these works: we examine the mechanics of the bargaining component of a regulatory penalty default regime, focus on specific implementations under a U.S. policy landscape, and discuss the advantages and disadvantages of this policy tool.

The rest of this paper is structured as follows. In Section 2, we provide a concise summary of the legal literature on penalty defaults for a computer science audience, and situate our contributions in related work. In Section 3, we introduce possible regulatory penalty default regimes for AI systems and discuss their merits and pitfalls. We conclude in Section 4.

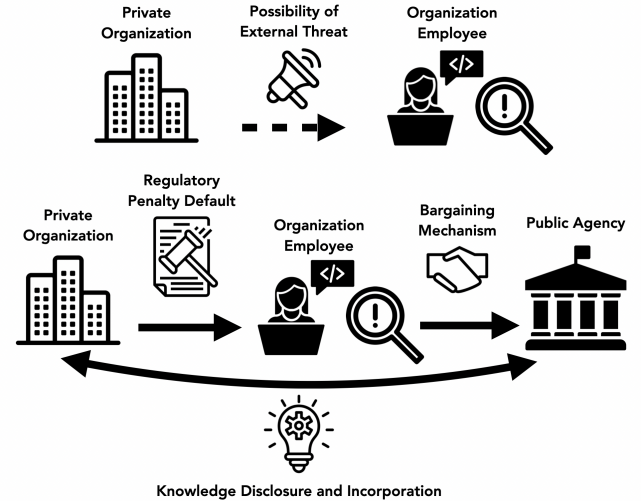

Figure 1: Illustration of incentives within different regulatory regimes. **Top (Self-Regulatory Regime)**: Under a self-regulatory regime, the possibility of external threats is an unreliable incentive to drive internal research into potential harms that AI systems might cause. **Bottom (Regulatory Penalty Default Regime)**: Under a regulatory penalty default regime, the regulatory penalty default induces private research into potential harms. This research can then be brought up as part of the bargaining process to a regulatory body, allowing for the transfer and generation of knowledge about building safer AI systems between private industry and public agencies.

2 BACKGROUND

In this section, we ground the notion of penalty default rules and regulatory penalty defaults in the legal literature. We also discuss related work.

2.1 Contracts and Penalty Defaults

In contract theory, a *complete* contract describes an outcome for the parties for every possible sequence of events. In contrast, *incomplete* contracts leave gaps in their specification that must be filled in after the fact. Due to a variety of information and resource constraints, most contracts used in practice are incomplete [44]. For example, contracting parties must deal with bounded rationality, transaction costs, legal costs from enforcement, and incomplete knowledge [10]. Research in contract theory often deals with the design and consequences of incomplete contracts–how to "fill in the gaps." One such method is a *default rule*. Default rules can fill the gaps in contracts by specifying what should occur unless certain conditions are met. This allows the contracting parties to focus on specific outcome where they should deviate from the default.

At first glance, one might suggest that a default rule should be designed to realize a mutually beneficial outcome. This "would have wanted" approach is a natural response to situations where the costs of contracting for specific contingencies lead to incompleteness. Such an approach may, e.g., arrive at an acceptable outcome while reducing transaction costs such as hiring lawyers or meeting at a venue. However, when contracts are incomplete for strategic

reasons, it is not clear that a mutually beneficial default is desirable. For example, a party with more information may take advantage of a desirable default to *withhold* information from the contract.

In their seminal paper on the subject, Ayres and Gertner [1] argue that penalty default rules are a way to manage these strategic incentives. The central idea is to introduce a default that is undesirable to the information-advantaged party. This creates an incentive for parties to disclose their private information and contract around the default. Ayres and Gertner [1] illustrate their proposal with the 19th century English contract case, Hadley v. Baxendale. In this case, Hadley contracted with Baxendale to ship a part necessary to run their mill. When the shipment was late, Hadley sued Baxendale to recover lost profits from closing the mill to wait for the part. Despite this, the court rules in Baxendale's favor because Hadley never disclosed the importance of the part to Baxendale. If Hadley had disclosed the importance of this part to his profit, Baxendale likely could have negotiated a higher price for shipping. Thus, a default rule that says that lost profits are unrecoverable creates an incentive for the recipient to disclose potential damages to the shipper. Ayres and Gertner [1] augment this informal argument with a game theoretic model and show that penalty defaults can induce information generating behavior. Bebchuk and Shavell [2] analyze a similar model based around communication cost. See Chapter 12 of De Geest [6] for a comprehensive overview of penalty default rules in contract theory[2].

2.2 Regulatory Penalty Defaults

Regulatory penalty defaults are a well-established tool in the collaborative governance literature [8, 16, 18, 37]. They are regulatory provisions or requirements that are undesirable, creating incentives for companies to cooperate or bargain around them [37]. Rather than to reprimand already-committed behaviors, the intent of regulatory penalty defaults is to induce preemptive private ordering in compliance with a regulatory standard, shifting the burden of information generation to regulated entities who have an information advantage. Because the penalty kicks in by default, regulators and regulated entities are situated as collaborators rather than as adversaries [37].

While regulatory penalty defaults intend to reduce information asymmetries as penalty default rules do in contracts, regulatory penalty defaults are not always the gap-filling *rules* that they are in contracts. In regulation, penalty defaults can often be exerted in more implicit ways. Selbst [37] notes that the "specter of regulation" can serve as a penalty default–just the threat of regulation can get private actors to get together and figure out how to regulate themselves; Garcia [8] notes that the cost of legal uncertainty for private entities that value legal certainty can serve as a penalty default; and Kaminski [16] suggests that the prescriptive broad standard that algorithms should not discriminate along with heavy fines whenever algorithms do discriminate is a regulatory penalty default–such a default would induce regulated entities to create standards around and mitigate discriminatory properties of algorithms.

As part of regulation, regulatory penalty default regimes have spurred the generation and disclosure of environmental standards, and the mitigation of environmental harms [18]. For example, as part of California's Proposition 65, the regulatory penalty default shifted the burden of proof onto corporations and other regulated entities to generate information about the toxicity of chemicals and show that specific use cases are safe. The penalty default in this case is the requirement of a warning label enforced by the threat of civil liability [18]. Corporations have bargained around this default through proposing and justifying standards to a designated regulatory body responsible for the determination of toxicity levels. As a result, more than 300 standards regarding toxic substances were jointly generated within a few months after the policy was enacted–more standards than the federal command-and-control regulations at the time [19].

2.3 Related Work

There has been related research on the role that penalty defaults can play in AI regulation. Selbst [37] considers regulatory mechanisms that could be incorporated to create incentives for good faith private sector participation within an algorithmic impact assessment (AIA) regime. In [37], Selbst also discusses the roots of impact assessment policies in the National Environmental Protection Act (NEPA), which Karkkainen [18] identifies as a regulatory penalty default. The author similarly suggests the consideration of step penalties in spurring private action under an AIA regime. In this work, we expand on a specific regime for autonomous vehicles (AVs) that draws from the regulatory penalty default model in NEPA.

In [16], Kaminski introduces the notion of binary governance for algorithmic decision-making, a governance structure that combines systemic governance along with an individual rights regime. Kaminski [16] frames the use of penalty defaults as one part of a comprehensive proposal to address the harms of algorithmic decision-making systems. Our work applies some of the ideas the author draws together as part of the collaborative governance toolkit more broadly, such as collaboration with a regulatory body in pursuing legal safe harbor and leveraging right of action as an incentive mechanism. In this work, we also discuss the mechanics of the bargaining component of a regulatory penalty default regime and focus on specific implementations under a U.S. policy landscape.

Regulatory approaches that aim to advance similar goals of information disclosure and harm mitigation are proposed and discussed in [4] and [21]. Both works present specific collaborative governance approaches that utilize private sector knowledge to promote information generation and harm mitigation. Cihon et al. [4] discusses the role of certification regimes in the reduction of information asymmetry for AI ethics, and Lu [21] proposes a disclosure requirement as part of the mandatory Securities and Exchange Commission (SEC) filings that public companies must complete. Both the certification and disclosure approaches aim to advance similar goals of providing incentives for information disclosure and mitigating harms. However, we focus on the incentives surrounding

[2]It is worth noting that there has been pushback against Ayres and Gertner [1]. Posner [31] argues that penalty defaults are not common and Maskin [25] critiques the theoretical analysis in Ayres and Gertner [1]. This argument is largely orthogonal to the claims in this paper. We focus on the regulatory applications of penalty defaults and their potential for AI regulation.

a collaborative governance approach, while most of the certification regimes discussed in [4] are self-regulatory, and the disclosure regime proposed in [21] follows more of a top-down regime.

Under both regimes, the burden of harm identification and evaluation largely fall on the governing body. Lu [21] provides specific recommendations on the disclosures that should be required. Under the certification regimes described in [4], the criteria would be determined and evaluated by the regulatory body. Moreover, many of the certification regimes discussed are voluntary. However, self-regulatory certification regimes may not actually create incentives for companies to subject their AI systems or organizational practices for review. For example, the Malta Digital Innovation Authority has yet to receive any applications for their certification program [4]. The aim of a successful regulatory penalty default regime, on the other hand, would be to shift the burden of information generation to private companies. Within a criteria-oriented regime, companies may operate under a model of legal compliance to avoid penalties that regulatory bodies might put in place. In the penalty default case, on the other hand, because penalties are already in place, companies have an incentive to present a generally compelling case to regulatory bodies to escape the penalty default.

3 REGULATORY PENALTY DEFAULT REGIMES

In this section, we discuss regulatory penalty default regimes for AI systems and discuss their merits and pitfalls. In Section 3.1, we consider the requirement of an input-transparent alternative introduced in the United States Filter Bubble Transparency Act (FBTA) as a regulatory penalty default for recommender systems. As part of the bargaining mechanism to escape the penalty default, we suggest the disclosure of companies' mitigation measures to reduce harms that come with the utilization of user-specified data, as well as oversight by the Federal Trade Commission (FTC). In Section 3.2, we propose two regulatory penalty default regimes for autonomous vehicles: one that builds on the notion of contact responsibility introduced in [48], and another that builds on burdensome impact assessment generation in environmental regulation [18] and the regulatory power of the National Highway Traffic Safety Administration (NHTSA) to grant exemptions. Finally, in Section 3.3, we discuss private right of action as a general regulatory penalty default for AI systems, as well as governing bodies that may mediate the bargaining process.

3.1 Input-Transparent Alternatives as a Regulatory Penalty Default Regime for Recommender Systems

In this section, we consider an approach to the design of regulatory penalty defaults that builds on Proposition 65's use of warnings in combination with the threat of civil liability [18]. We propose a regulatory approach that combines a broad standard of potentially harmful algorithms with a penalty default. Unlike the warnings used in Proposition 65, we suggest that the "input-transparent" requirement from the proposed FBTA may function as an effective penalty default to spur investment in measuring, disclosing and mitigating harms from recommender systems.

Recommender systems are data-driven algorithms that choose, order, and present content to users on platforms. They have many applications, such as video-streaming platforms, social media, news, and marketplaces. There are growing concerns about recommender systems, such as their role in addictive tendencies on video-streaming platforms and the proliferation of misinformation on social media [28]. As a result of a 2019 congressional hearing that discussed these claims, the United States senate introduced the FBTA in June of 2021 [39]. The act requires that companies of a certain size provide notice and an "input-transparent" alternative when platforms use data-driven, opaque algorithms to show content.

The proposed policy will provide users some choice to opt-out of algorithmic recommendation. However, it will not spur private actors to, e.g., develop standards for fair or beneficial recommendation systems. An alternative would be to modify the FBTA to fit within a regulatory penalty default regime. In this regime, one could specify by default that input-transparent alternatives are required and, potentially, that any algorithmic recommendation be strictly opt-in. This requirement would constitute a significant penalty, because the input-opaque version of the recommender system drives the company's bottom line and default options on the internet are frequently unchanged.

To complement this default, a penalty default regime specifies a pathway to bargain around it. In this case, one could require companies to justify that their recommender system is not harmful. The FTC's goal of consumer protection against deception and its experience monitoring recommender systems through privacy consent decrees [7] make it a good match for this role. The bargaining process under this regime would involve oversight mechanisms similar to those that the the commission has administered for privacy oversight in the past. Moreover, the FTC has additionally been designated with the enforcement of the recently proposed Algorithmic Accountability Act of 2022, overseeing the creation of structured guidelines for corporate assessment and reporting and acting as a public repository for impact assessments [47].

With addiction and misinformation recognized as top concerns in recommender systems, the FTC could require extensive internal reviews for those harms that are coupled with independent assessments of companies' progress towards those standards. These processes can generate information about which aspects of a recommender system can be modified to make the system less harmful to users. Furthermore, a company's desire to create a convincing case to the FTC may lead to the development of new recommendation approaches that are transparent and beneficial, as we have seen in companies' "beyond compliance" efforts as part of consent decrees. Such a regime for recommender systems could spur the knowledge generation process in the creation of those structured guidelines. For example, if the FTC recognizes the effectiveness of a particular approach included in a company's bargaining packet, the regulatory body can include that approach as part of its structured guidelines for corporate assessment. On the other hand, if the FTC notes the ineffectiveness of a particular approach, the regulatory body can include that approach as part of its case studies. Importantly, this regime places the onus of knowledge generation and justification on the company, rather than on the FTC. This promises greater levels of transparency for regulators into the functionality of recommender systems.

This disclosure and transparency is an important regulatory goal. Stray [43] notes that Facebook and YouTube have tweaked the metrics of their recommender systems to increase user well-being under their respective definitions. However, the companies have not yet published the results of their studies. This hinders the regulatory goal to embed societal values in algorithmic governance [33]. Corporate disclosures of contextual information to a regulatory body as part of the bargaining process creates an opportunity for regulatory bodies to generate public reports that support a society-in-the-loop model of algorithmic governance.

While this approach may help create knowledge about and standards for safer recommender systems, companies still have a lot of leeway. For example, they could design the input-transparent alternative to be lower quality so that users are more likely to opt-in to algorithmic recommendations. However, the intentional deterrence of users to the input-transparent version of recommender systems is a downside that plagues both our penalty default regime and the proposed Filter Bubble Transparency Act. Additionally, this approach, like the FBTA itself, relies on an effective and enforceable definition of input-transparent alternatives. To manage these pitfalls, an effective implementation of this idea will rely on efficient approaches to external monitoring more generally.

3.2 Regulatory Penalty Default Regimes for Autonomous Vehicles

Under contact responsibility, AV companies internalize the costs for all accidents in which the AV is implicated [48]. Wansley [48] notes that public records of AV crashes demonstrate the ability of AVs to avoid causing crashes, but not the ability to avoid "plausibly preventable crashes caused by human error". A contact responsibility standard would mean that an AV company cannot simply evade responsibility for the costs of a crash by relying on existing liability and negligence mechanisms, such as pointing to negligent human driving behavior (e.g., drunk driving). This creates an incentive for AV companies to invest in engineering infrastructure to anticipate human driving behaviors and produce vehicles that drive defensively [48]. Wansley [48]'s system, which holds AV systems to a different standard of legal liability, can also fit into a regulatory penalty default. Contact responsibility, or the level of contact responsibility, can induce private ordering around the creation of safer AVs. A bargaining process to reduce contact liability can create an additional incentive to disclose best practices for safe AV design. For example, companies could face more liability (e.g., triple damages) by default. The path to reduce this liability relies on bargaining with a governing agency (e.g., NHTSA) to justify safety measures in place and disclose relevant metrics.

An alternative regulatory penalty default regime for AVs draws from NHTSA's regulatory power to grant exemptions and NEPA's regulatory penalty default of a burdensome impact statement [18]. Under the Safety Act [46], NHTSA has the authority to grant exemptions from Federal Motor Vehicle Safety Standards (FMVSS). The FMVSS includes specific rules for conventional vehicles (CVs), e.g., requirements about an unobstructed view of the rear of a vehicle from the driver's seat. Nuro, a California-based AV company, has worked with NHTSA to bargain around and justify exemptions from the FMVSS [30]. As part of Nuro's application process for an exemption from certain FMVSS, NHTSA evaluated the FMVSS-exempt AV in terms of whether it would achieve an equivalent level of safety as a compliant version of the AV would achieve. For instance, in NHTSA's assessment of Nuro's petition for exemption, NHTSA considered differences in safety levels between an exempt and a compliant version of Nuro's R2X vehicle: "The question of whether an exemption would lower the safety of an exempt version of the R2X as compared to a compliant version of the vehicle turns on the very limited differences between those two versions of the R2X, which are only that the exempted R2X would not comply with the certain requirements described in this notice" [30]. In approving Nuro's petition, NHTSA also sought input from industry and academic stakeholders and required Nuro to provide a report every 90 days that includes material changes made to the Automated Driving System software, among other disclosure requirements [30].

This process functioned similarly to how a bargaining process might be undertaken under a penalty default regime. The burden of proof rested on Nuro to present a compelling case that the company's exempt AVs would be just as safe as vehicles compliant with FMVSS. On the regulatory end, NHTSA evaluated many of Nuro's AV features under an "equivalent overall safety" standard, asking whether an exempt version of the vehicle would lead to reductions in vehicle safety. However, holding AVs to an "equivalent overall safety" standard to CVs that are compliant with FMVSS may not create incentives for company documentation and disclosure of features integral to building the infrastructure for AVs to drive more safely than humans.

At the same time, because granular FMVSS rules were written for CVs (a rearview mirror positioned in a certain way for an AV without a human operator would be quite vestigial), it is likely difficult to produce an AV that complies with all of them. Furthermore, it is burdensome, from a company perspective, to identify and justify exemptions from all of the human-operator specific FMVSS. We note "human-operator specific" because some FMVSS rules still apply to AVs. For example, standards that deal with tire and material flammability ensure that bodies of AVs are designed to be safe in the way that the body of CVs are. At present, a vehicle manufacturer must only show that the exempt version does not lead to any reductions in safety. However, the fundamental question of interest for AV regulation is not how a vehicle is designed so that a human can safely operate it. Rather, it is how AV infrastructure is researched and engineered to manage risk from accidents.

As an alternative, the requirement of burdensome identification and justification for exemptions from human operator-specific FMVSS might serve as a penalty default that AV companies can bargain around by disclosing harms, challenges, and mitigation measures specific to AV design. Such an arrangement could operate in tandem to contact responsibility as introduced in Wansley [48]. With the proposed contact responsibility regime providing incentives to invest in safety, an FMVSS penalty default may serve as a way to provide incentives for the disclosure of AV-specific engineering challenges and preemptive mitigation measures towards defensive driving. NHTSA has already required AV companies to disclose specific requirements as part of the exemption from FMVSS [30]. Instead of fulfilling those requirements to meet compliance, a regulatory penalty default might instead provide incentives for AV

companies to make a compelling case to escape the default and identify harms outside of NHTSA's requirements.

This approach is similar to the ways that NEPA operates as a penalty default. Under NEPA, the Environmental Protection Agency (EPA) requires federal agencies to produce an environmental impact statement (EIS), a statement detailing the expected environmental impacts of a proposed plan of action and possible alternatives to the course of action, when a project is expected to incur a significant cost to the environment [18]. Federal agencies have bargained around this requirement by keeping the expected environmental impacts of their projects below impact thresholds through "adding mitigation measures" or "redefining projects" through an environmental assessment (EA) such that the projects stay below the threshold after which an EIS would be required [18].

For AV regulation, an analogous regulatory penalty default might allow an AV company to bargain around the identification of and justification for exemption from human operator-specific FMVSS by working with NHTSA to detail measures regarding how the company expects to minimize the externalities of its vehicles as part of an impact assessment and project management plan. This exercise of proposing a project management plan may also encourage the imagination of possible futures and of overcoming harms in the way that is recognized as important in AI design. Boyarskaya et al. [3] claims that failures to anticipate harms in AI systems are often the result of the failures to imagine and ask how to mitigate harms. Boyarskaya et al. [3] additionally emphasizes the importance of context-aware frameworks of harm in broadening the range of considerations, given the range of ways in which AI systems can cause harm. While certain externalities of AVs are uniquely visible and high-profile (crashes and deaths), the underlying errors that can give rise to them are endless–e.g., not "seeing" a truck against a bright blue sky [49], or not adequately anticipating how soon the driver behind it will brake [42]. Moreover, crashes are not the only externalities that should be considered in the deployment of AVs. Project management plans may also include company considerations that AVs are deployed to optimize traffic flow [9].

With NHTSA's current close monitoring and engagement with AV companies, regulatory support for a penalty default regime with NHTSA as the bargaining regulatory body could be possible. Drawbacks to an FMVSS penalty default regime with an impact assessment, however, include path dependency, legal ambiguity, and trade secrecy. While burdensome, there is now precedence for identifying and justifying for exemptions from human operator-specific FMVSS. Then, identifying and justifying for exemptions from human operator-specific FMVSS may not be burdensome enough for companies to risk the legal ambiguity of whether their AV-specific infrastructure impact assessment is sufficient enough to escape the penalty default. Additionally, presenting the impact assessment would release information specific to the company's AV vehicle and technologies, risking the release of trade secrets. However, company disclosures would be facilitated by regulatory bodies. Regulatory bodies can serve as a medium of information disclosure to the public while utilizing that information to design better regulations. For example, Waymo, another autonomous vehicle (AV) company spun out of Google, is disclosing information to the California Department of Motor Vehicles (CA DMV) as part of an AV pilot program. The company sued successfully to keep aspects of its emails between the DMV as well as technical details about its vehicles retracted from the public for a period of time, but available to the CA DMV [12]. Ultimately, bargaining around ex ante requirements may create incentives for safer AV infrastructure design and preemptive disclosure of those design decisions.

3.3 Private Right of Action as a Penalty Default Regime

Another potential for a regulatory penalty default regime leverages right of action as a background threat of regulation. This was briefly discussed in [16]. Here, we consider this proposal in depth and discuss how it could build on existing legal precedent in the United States.

Private right of action is the right of individual citizens to bring a lawsuit directly against a company. It is an important legal mechanism in the United States that has already structured corporate incentives in privacy legislation [35]. Scholz [35] notes that, where privacy regulation is solely publicly enforced, typically by state attorneys general and the FTC, existing regulatory loopholes are easily exacerbated. Public enforcement cannot take into account every wrong. With private right of action, every potential wrongdoing can become the subject of litigation, creating incentives for companies and agencies to comply with the law in every consumer interaction. Private right of action has also been effective at surfacing violations of antitrust, with almost half of antitrust cases discovered by private attorneys [20, 35].

In the context of AI, private right of action has had a pronounced effect for the regulation of facial processing technologies (FPT) in the United States. FPT are AI systems that identify faces in images and use that information in downstream predictions. Several states have biometric privacy laws. However, the state of Illinois' has led to the clearest response from companies. Notably, their law, the Biometric Information Privacy Act (BIPA) is the only one that provides a private right of action for the nonconsensual processing of biometric information [11]. As a result, BIPA has meaningfully influenced company decisions in their deployment of FPT. For example, Clearview AI announced that they would end their contracts with all non-enforcement entities in Illinois [22]. Google disabled its Arts & Culture face feature in Illinois [24], and Sony AI chose not to sell their robot dog that utilizes FPT in Illinois [40].

It is not our goal to comment on the benefits of BIPA directly. Arguably, the response of companies to pull their products entirely is not ideal. Instead, our goal is to highlight the fact that private right of action has a demonstrable effect on the incentives of technology companies. This suggests that private right of action may be effective as a regulatory penalty default. In order to include it in a regulatory penalty default regime, it would need to be combined with a bargaining method to avoid the penalty. In this case, the legislature would empower a regulatory body to approve mitigation measures for harms technologies might cause, on a case by case basis. Once approved, the technology would still be subject to lawsuits from, e.g., a state's attorney general, as is the case for biometric privacy laws in Washington and Texas, or third-party commissioners, but may not be subject to lawsuits from private individuals.

The standard required to avoid the penalty default should be based on the clear disclosure of a potential harm that the company wants to limit liability for and the suitable description of a mitigation strategy for the harm. Companies could propose, e.g., a combination of internal and external audits along with regular disclosures that would allow regulators to monitor the effectiveness of mitigation efforts. A useful feature of this regulatory penalty default is that it is generic. Private right of action can easily be applied to a variety of AI systems, not just FPT. For example, this approach could also be applied to medical devices or autonomous vehicles, for which domain-specific regulatory bodies like the Food and Drug Administration (FDA) and the National Highway Traffic Safety Administration (NHTSA) exist to manage the bargaining process. In the case of BIPA, the introduction of a bargaining option may create incentives for companies to develop FPT methods that effectively address lawmaker's concerns, instead of pulling out of the market.

There are also several drawbacks to private right of action as a penalty default. First, it may be important to allow individuals to contest AI systems in order to exercise individual rights and address the dignatary and justificatory harms that the development and deployment of AI systems may cause [17]. If private right of action is used as a penalty default, then it is treated as a policy tool to structure corporate incentives, rather than one that directly redresses individual harms. There is potential, however, for commissions such as the United States Equal Employment Opportunity Commission (U.S. EEOC)[3] to fill this regulatory role and allow for meaningful recourse from harms resulting from AI hiring tools. For example, private right of action has not entirely sufficed for the actual governance of AI hiring algorithms. Private right of action is typically the only avenue for legal recourse in discriminatory hiring cases. This puts the burden of proof to show discrimination through disparate treatment and disparate impact on plaintiffs [32]. Plaintiffs, especially those unfairly deprived of job opportunities by discriminatory hiring algorithms, may not have the resources needed to demonstrate discrimination in a legal sense. On the other hand, the EEOC, through its Artificial Intelligence and and Algorithmic Fairness Initiative [5], may be in a better position to pursue meaningful action against companies that develop unfair or discriminatory hiring algorithms.

A second drawback to this penalty default approach is the challenge of determining the appropriate regulatory body to assess claims for AI systems across different domains. An effective regulatory penalty default regime relies on an effective regulatory body to assess proposals and determine whether disclosed harms that arise from a company's products are appropriately mitigated. In the cases considered above, the FTC and NHTSA were natural options. In general, for AI systems that already operate in domains that are already regulated, existing regulatory bodies for those areas could fill such a role. However, for generic AI systems, especially AI systems that introduce new domain areas, it is not clear how to allocate this responsibility.

[3]Thanks to Manish Raghavan for this point.

4 CONCLUSION

Regulatory penalty default regimes are a potentially useful tool for the regulation of AI systems. Under the current self-regulatory approach to AI harms in the United States, the possible threat of negative publicity drives a reactive approach to harm mitigation. This possible threat assigns an intangible value to private sector employees who study and mitigate the harms that their AI systems may cause, and encourages a reactive approach to harm mitigation. This paper explores how regulatory penalty default regimes can create incentives for the preemptive information generation about and the mitigation of harms in the private sector. The disclosure of harms and proposed mitigation measures are crucial to bargaining around the defaults, which, if successful on the company's end, can lessen the severity of the penalty. This gives private sector employees who identify and mitigate harms tools to advocate for their work internally. This is because research that supports successful bargaining justifications leads to concrete reductions in penalties and substantiates the value their work brings to their company.

At the same time, regulatory penalty default regimes also have limits and drawbacks. The approach relies on an effective regulatory body to manage the bargaining process. The potential abuse of regulatory power points to the need for opportunities for other affected stakeholders and community members to voice concerns and meaningfully influence regulation. This could be addressed, for example, as part of a regulatory penalty default regime that is similar to NHTSA's invitation for stakeholders to provide comment on AV companies' documentation [30]. Another approach to this concern could be similar to California's Community Air Protection Act, which brings additional resources to communities most impacted by air pollution and has led to the creation of Community Steering Committees [41][4]. As part of regulatory penalty default regimes for AI systems, disclosures in bargaining processes could similarly involve affected stakeholders.

The core contributions of this paper are a synthesis of the penalty default literature for a computer science audience and consideration of how regulatory penalty default regimes may be applied across AI domains under a U.S. policy landscape. It is important to note that regulatory penalty default regimes are just one approach at addressing the issue of harm mitigation and disclosures for AI systems. The use of regulatory penalty defaults does not necessarily preclude the use of other regulatory tools, such as certification or disclosure. In fact, regulatory penalty defaults can be used to shape incentives in achieving these regulatory ends. To our knowledge, while regulatory penalty defaults are well-established in the collaborative governance literature, regulatory penalty default regimes for the oversight of AI systems have not yet been implemented. Future work should consider budget and time constraints for regulatory agencies that can oversee potential bargaining mechanisms and formalize cost considerations for private organizations. Overall, regulatory penalty default regimes are not a panacea, but they are a useful tool to spur private generation and disclosure of knowledge for the governance of AI systems.

[4]Thanks to Christina Chen for this point.

5 ACKNOWLEDGEMENTS

We thank Lawrence McCray for helpful clarifying questions throughout this research project, as well as for identifying regulatory bodies of interest. Additionally, we thank Andreas Haupt, Andi Peng, and Stephen Casper for valuable discussions. Finally, we thank the reviewers for their helpful feedback.

REFERENCES

[1] Ian Ayres and Robert Gertner. 1989. Filling gaps in incomplete contracts: An economic theory of default rules. *The Yale Law Journal* 99, 1 (1989), 87–130.
[2] Lucian A Bebchuk and Steven Shavell. 1991. Information and the scope of liability for breach of contract: The rule of Hadley v. Baxendale.
[3] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. *arXiv preprint arXiv:2011.13416* (2020).
[4] Peter Cihon, Moritz J Kleinaltenkamp, Jonas Schuett, and Seth D Baum. 2021. AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries. *IEEE Transactions on Technology and Society* (2021).
[5] U.S. Equal Employment Opportunity Commission. 2021. EEOC Launches Initiative on Artificial Intelligence and Algorithmic Fairness. (2021). https://www.eeoc.gov/newsroom/eeoc-launches-initiative-artificial-intelligence-and-algorithmic-fairness
[6] Gerrit De Geest. 2011. *Contract law and economics*. Vol. 6. Edward Elgar Publishing.
[7] FTC. 2019. Federal Trade Commission and People of the State of New York v. Google LLC and Youtube LLC. (2019). https://www.ftc.gov/system/files/documents/cases/172_3083_youtube_coppa_consent_order.pdf
[8] Kristelia A Garcia. 2014. Penalty default licenses: A case for uncertainty. *NYUL Rev.* 89 (2014), 1117.
[9] Thomas Krendl Gilbert, Sarah Dean, Nathan Lambert, Tom Zick, and Aaron Snoswell. 2022. Reward Reports for Reinforcement Learning. *arXiv preprint arXiv:2204.10817* (2022).
[10] Dylan Hadfield-Menell and Gillian K Hadfield. 2019. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 417–422.
[11] Woodrow Hartzog. 2020. BIPA: The Most Important Biometric Privacy Law in the US? *Regulating Biometrics: Global Approaches and Urgent Questions, ed. Amba Kak (AI Now 2020)* (2020), 96–103.
[12] Andrew J. Hawkins. 2022. Waymo sues California DMV to keep driverless crash data under wraps. (2022). https://www.theverge.com/2022/1/28/22906513/waymo-lawsuit-california-dmv-crash-data-foia
[13] Donald T Hornstein. 2004. Complexity theory, adaptation, and administrative law. *Duke LJ* 54 (2004), 913.
[14] Elisa Jillson. 2021. Aiming for truth, fairness, and equity in your company's use of AI. *Federal Trade Commission* (2021). https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai
[15] Margot E Kaminski. 2015. When the default is no penalty: negotiating privacy at the NTIA. *Denv. L. Rev.* 93 (2015), 925.
[16] Margot E Kaminski. 2018. Binary governance: Lessons from the GDPR's approach to algorithmic accountability. *S. Cal. L. Rev.* 92 (2018), 1529.
[17] Margot E Kaminski and Jennifer M Urban. 2021. The right to contest AI. *Columbia Law Review* 121, 7 (2021), 1957–2048.
[18] Bradley C Karkkainen. 2005. Information-forcing environmental regulation. *Fla. St. UL Rev.* 33 (2005), 861.
[19] Bradley C Karkkainen. 2008. Framing Rules: Breaking the Information Bottleneck. *NYU Envtl. LJ* 17 (2008), 75.
[20] Robert H Lande and Joshua P Davis. 2007. Benefits from private antitrust enforcement: An analysis of forty cases. *USFL Rev.* 42 (2007), 879.
[21] Sylvia Lu. 2020. Algorithmic Opacity, Private Accountability, and Corporate Social Disclosure in the Age of Artificial Intelligence. *Vand. J. Ent. & Tech. L.* 23 (2020), 99.
[22] Ryan Mac, Caroline Haskins, and Logan McDonald. 2020. Clearview AI Has Promised To Cancel All Relationships With Private Companies. *Buzzfeed News* (2020). https://www.buzzfeednews.com/article/ryanmac/clearview-ai-no-facial-recognition-private-companies
[23] International Business Machines. 2022. Our foundational properties for AI ethics. *IBM* (2022). https://perma.cc/2ZSV-LFLE
[24] Ally Marotti. 2018. Google's art selfies aren't available in Illinois. Here's why. *Chicago Tribune* (2018). https://www.chicagotribune.com/business/ct-biz-google-art-selfies-20180116-story.html
[25] Eric Maskin. 2005. On the rationale for penalty default rules. *Fla. St. UL Rev.* 33 (2005), 557.
[26] Lachlan McCalman, Daniel Steinberg, Grace Abuhamad, Marc-Etienne Brunet, Robert C Williamson, and Richard Zemel. 2022. Assessing AI Fairness in Finance. *Computer* 55, 1 (2022), 94–97.
[27] Microsoft. 2022. Responsible AI Principles from Microsoft. *Microsoft* (2022). https://perma.cc/2G6B-UYG3
[28] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *Ai & Society* 35, 4 (2020), 957–967.
[29] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. 2021. Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY* (2021), 1–13.
[30] NHTSA. 2019. Nuro, Inc.; Grant of Temporary Exemption for a Low-Speed Vehicle with an Automated Driving System. (2019).
[31] Eric A Posner. 2005. There are no penalty default rules in contract law. *Fla. St. UL Rev.* 33 (2005), 563.
[32] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.
[33] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.
[34] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 7 (April 2021), 23 pages. https://doi.org/10.1145/3449081
[35] Lauren Henry Scholz. 2021. The Significance of Private Rights of Action in Privacy Law. *William & Mary Law Review, Forthcoming* (2021).
[36] Andrew D Selbst. 2020. Negligence and AI's human users. *BUL Rev.* 100 (2020), 1315.
[37] Andrew D Selbst. 2021. An Institutional View Of Algorithmic Impact Assessments. (2021).
[38] Andrew D. Selbst, Suresh Venkatasubramanian, and I. Elizabeth Kumar. 2021. The Legal Construction of Black Boxes. In *We Robot 2021*.
[39] Senate - Commerce, Science, and Transportation. 2021. Filter Bubble Transparency Act. (2021).
[40] SonyAI. 2018. Why Is aibo Not for Sale in Illinois? *SonyAI* (2018). https://www.sony.com/electronics/support/articles/00202844
[41] South Coast AQMD. 2020. AB 617 Community Air Plan Community Steering Committee Charter. (2020). https://www.aqmd.gov/docs/default-source/ab-617-ab-134/steering-committees/southeast-los-angeles/charter-feb-2020.pdf?sfvrsn=8
[42] Jack Stewart. 2018. Why People Keep Rear-Ending Self-Driving Cars. (2018). https://www.wired.com/story/self-driving-car-crashes-rear-endings-why-charts-statistics/
[43] Jonathan Stray. 2020. Aligning AI Optimization to Community Well-Being. *International Journal of Community Well-Being* 3, 4 (2020), 443–463.
[44] Jean Tirole. 1999. Incomplete contracts: Where do we stand? *Econometrica* 67, 4 (1999), 741–781.
[45] Charlotte A Tschider. 2020. Medical Device Artificial Intelligence: The New Tort Frontier. *BYU L. Rev.* 46 (2020), 1551.
[46] United States Congress. 1966. National Traffic and Motor Vehicle Safety Act. (1966).
[47] United States Senate. 2022. Algorithmic Accountability Act. (2022).
[48] Matthew Wansley. 2021. The End of Accidents. (2021).
[49] Danny Yadron and Dan Tynan. 2016. Tesla driver dies in first fatal crash while using autopilot mode. (2016). https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk