Regulating Algorithms: What and When

Talia Gillis Scott Nelson Jann Spiess

May 31, 2025

Abstract

The regulation of algorithmic decisions, ranging from advanced credit scoring to employment screening, presents unique challenges and novel opportunities for achieving regulatory goals. We propose a framework for algorithmic regulation that emphasizes the importance of temporal stages in the regulatory pipeline: ex-ante (pre-training), ex-interim (post-training but pre-deployment), and ex-post (post-deployment). Regulators can choose both the pipeline stage targeted by the legal rule ("rule timing") and the stage at which compliance is assessed ("scrutiny timing"). We situate emerging and proposed AI regulations within this framework and analyze the tradeoffs between different regulatory regimes. We highlight how ex-interim rules offer a unique opportunity in algorithmic settings compared to the rigidity of ex-ante rules or the bluntness of ex-post rules and explore the considerations that guide whether regulators might scrutinize ex-interim rules before or after deployment. We conclude this chapter by outlining an agenda for developing effective tools to regulate AI.

Authors are listed in alphabetical order. We thank Laura Blattner for earlier input and discussions, Haggai Porat for insightful comments, and Ritha Sarf for excellent research assistance.

I. Introduction

In many high-stakes domains of algorithmic decision-making, those who develop and deploy complex AI systems may have different objectives from the regulators overseeing them. Firms deploying algorithmic systems may prioritize profitability or other private benefits, while regulators may have additional or differing concerns, such as the distributional effects of algorithmic decisions and their impact on safety and market stability. These differences in preferences, compounded by information and technological asymmetries and the costs of regulation, present a fundamental challenge—how should algorithmic decisions be regulated?

This chapter examines regulatory approaches available for overseeing algorithmic decision-making, with a particular focus on the temporal design of rules and their evaluation. Our goal is to illuminate the trade-offs between different approaches and highlight the unique opportunities afforded by algorithmic settings compared to more traditional decision-making contexts.

In our framework, a firm develops and deploys a policy for automated decision-making—such as a lender creating a loan-underwriting model to maximize profits—while a regulator considers additional societal objectives like fairness, safety, or systemic risk. The regulator aims to design a regulatory regime that addresses the misalignment of preferences while accounting for information and cost constraints. The interaction between regulator and lender plays out around two key phases of algorithmic development and deployment. In the training phase, some initial data become available, and the firm uses these data to implement its algorithm. Next, in the deployment phase, the resulting algorithmic policy is applied for decision-making on new data—such as real-time loan applications—the particulars of which might not have been fully knowable during training.

We argue that in algorithmic settings, there is a distinctive opportunity to design regulation around these two phases of model training and deployment. We highlight two temporal dimensions of regulation. The first dimension concerns the stage of the algorithmic pipeline targeted by a legal rule: *ex-ante* (before the training phase), *ex-interim* (after training but before deployment), or *ex-post* (after deployment). *Ex-ante* rules put coarse constraints on the models that can be trained, even before the data become available. For instance, such rules might limit the permissible inputs (e.g., prohibiting the use of protected class variables) or restrict the models to certain types, such as simple or self-interpretable models. These restrictions can shape the model design space before any training data are observed. *Ex-interim* rules apply after the model has been trained but before deployment, and constrain the models adopted by the firm. For example, one way these rules can operate is by allowing regulators to inspect and test models prior to those models' deployment. Finally, *ex-post* rules target the deployment stage, where regulators evaluate the outcomes produced by the deployed model. These outcomes reflect not only the algorithmic policy but also the particulars of the real-world environment—deployment state—that are only knowable at this stage. For example, regulators may investigate disparities in realized lending decisions, which arise from the interaction between the deployed algorithmic policy and the actual distribution of applicants.

The second temporal dimension concerns the timing of evaluating compliance with a rule. A legal rule can be scrutinized at a stage later than the one it targets. For example, an ex-ante rule might be scrutinized not only during the ex-ante stage but also at the ex-interim or ex-post stages following model development and/or deployment. Similarly, ex-interim rules may be scrutinized either during the ex-interim stage or later at the ex-post stage. By explicitly considering these two temporal dimensions—the timing of rules and the timing of

scrutiny—this chapter maps the space of possible regulatory regimes and explores the trade-offs associated with different approaches.

We leverage this framework to classify emerging approaches in AI regulation. Specifically, we develop a temporal taxonomy that maps regulatory approaches according to both the stage of the algorithmic decision-making pipeline targeted by the rule and the timing at which compliance with the rule is assessed. This taxonomy allows us to systematically examine how current regulations, proposed frameworks, and policy discussions address the full range of regulatory options.

Our main insight is that ex-interim rules represent a potentially powerful tool for regulating algorithmic policies. Unlike ex-ante interventions, ex-interim rules are less rigid because they account for the specific models developed based on the training data. Compared to ex-post interventions, ex-interim rules are less blunt since they more directly target the *conduct* of the firm, as they do not depend on the (potentially random) realization of outcomes that the firm has no direct control over. Crucially, ex-interim rules, particularly when scrutinized before deployment, can help prevent the materialization of harmful outcomes. This offers an important opportunity in algorithmic settings compared to more traditional decision-making contexts. For example, in decision-making based on human discretion, there is often no ex-interim stage where a policy is fully described or describable prior to deployment, leaving only rigid ex-ante or blunt ex-post regulation as feasible options. When decisions are taken by algorithms, on the other hand, they are mathematical objects that can be described, tested, and evaluated under controlled conditions even before they are applied.

To illustrate the potential value of ex-interim scrutiny, consider a financial regulator overseeing algorithmic credit decisions by a lender. The regulator may, for example, aim to ensure that not too many loans are extended to risky borrowers. However, the number of risky loans issued and the number of actual defaults depend not only on the lender's algorithmic policy, but also on factors beyond the lender's control, such as which borrowers apply and how economic conditions affect repayment behavior. While the algorithm is under the lender's control, some of these external factors are not. When regulation is designed to target conduct rather than outcomes alone, scrutinizing the lender only based on realized outcomes may be inefficient, as it risks penalizing the lender for bad luck rather than for unreasonable decisions. Evaluating whether the algorithmic policy itself was reasonable based on information available at the time of development may lead to a more targeted and effective regulatory policy. Furthermore, such scrutiny is possible even before the lending rule is deployed, allowing regulators to intervene before harm is done.

The desirability of ex-interim rules, however, depends crucially on the cost of scrutinizing compliance by the firm at each stage. If scrutiny of compliance with ex-interim rules is more resource-intensive than scrutinizing ex-post rules—for instance, because ex-post rules leverage easily measurable outcomes, while scrutinizing ex-interim rules may require accessing the training data and evaluating a highly complex algorithmic process—then regulation based on ex-interim rules may be less attractive for a resource-constrained regulator. Despite these caveats, we illustrate in a stylized example that ex-interim scrutiny can have value even in cases when it has higher cost, since targeting the firm's conduct directly allows the regulator to align preferences in a more precise way than overly static or overly noisy ex-ante and ex-post rules can. The effectiveness of ex-interim rules also depends on factors such as the relationship between the training signal and the deployment state. For example, if the training data used by a lender closely resembles the real-world distribution of borrowers, the outcomes of the deployment stage are more likely to reflect the decisions made during the training phase. In such cases, ex-post

rules are less likely to target noise or unexpected factors unknown to the developer at the training stage.

Our chapter contributes to the interdisciplinary literature on regulatory design as it relates to AI, spanning economics, law, political science, computer science, and data science more broadly. Recent work examines the design of regulations for algorithmic decision-making and the trade-offs faced by regulators when aligning firm incentives with broader societal objectives (Rambachan et al., 2020; Guerreiro, Rebelo, and Teles, 2023; Cowgill and Tucker, 2020). The work in Kleinberg et al. (2018) and Kleinberg et al. (2020) emphasizes the opportunities for ex-interim legal rules in algorithmic settings, where algorithmic policies are specified before deployment. Our work extends these insights by emphasizing additional temporal dimensions in the regulatory pipeline. Specifically, we study the trade-offs between regulating algorithms before and after deployment, building on our previous work introducing the concept of algorithmic explainers (Blattner, Nelson, and Spiess, 2024) and stress testing (Gillis and Spiess, 2019) at the ex-interim stage.[1]

Other related work examines current and emerging AI regulatory frameworks across countries (Comunale and Manera, 2024) and specific regulatory tools, such as licensing and auditing (Guha et al., 2023; Anderljung et al., 2023; Hadfield and Clark, 2023), which we consider within our temporal framework. Our work also builds on literature documenting and considering the various AI risks, such as existential risks (Jones, 2023; Acemoglu, 2021), privacy (e.g., Goldfarb and Tucker, 2012; Acquisti, Taylor, and Wagman, 2016), and fairness (e.g., Hardt, Price, and Srebro, 2016), and the tensions between innovation and addressing potential risks (Callander and Li, 2024; Agrawal, Gans, and Goldfarb, 2023). By relating these different policy considerations to the regulatory timing in algorithmic pipelines, we demonstrate the trade-offs associated with legal rule timing. While we highlight some representative examples here, this list is not exhaustive, as the literature on AI regulation continues to grow rapidly.

Our analysis also relates to the classic law and economics literature on optimal regulatory design, which discusses the choice of regulatory restriction types (Shavell, 2018; Kaplow, 2013), the design of legal enforcement mechanisms (Shavell, 1984) and the impact of enforcement costs on optimal legal rules (Shavell, 1993; Kaplow and Shavell, 1994).

The remainder of this chapter proceeds as follows. Section II provides a high-level temporal framework of algorithmic decision-making and regulation. Section III presents our taxonomy of regulatory approaches within that framework, distinguishing between when rules apply in the algorithmic pipeline and when compliance with those rules is assessed. Section IV builds on this taxonomy by formally modeling regulation as a game between a firm deploying an algorithm and a regulator overseeing its use, and illustrates key considerations through a simple, stylized example. Section V highlights the model's key insights and situates them within related work, emphasizing the unique opportunity that algorithmic regulation provides to scrutinize decisions before deployment and the materialization of harm. This section also discusses some important limitations of our framework. Finally, Section VI proposes some directions for future research.

[1] Another related literature has studied the welfare implications of personalized pricing when firms use "big data" to price discriminate (Bergemann, Brooks, and Morris, 2015; Porat, 2022; Dubé and Misra, 2023; Ali, Lewis, and Vasserman, 2023; Bar-Gill, Sunstein, and Talgam-Cohen, 2023; Rhodes and Zhou, 2024). While this literature has largely focused on the policy question of how and whether to restrict what data firms can use, we show that there are additional regulatory possibilities that, while allowing broad access to rich consumer data, regulate the pricing models that use these data.

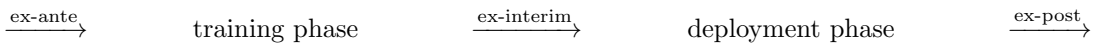

Figure 1: Stylized temporal structure of the algorithm pipeline and its regulation

II. A Temporal Framework for the Regulation of Algorithms

We consider a high-level framework of algorithmic deployment and its regulation. This multi-stage framework covers the interaction between a firm deploying an algorithm and a regulator overseeing its use. The firm chooses an algorithmic policy. For example, the firm could be a lender that designs an algorithm to make credit approval decisions. The regulator puts constraints or imposes penalties on the algorithm and its implications. For example, a financial regulator may restrict the variables that are permissible for use in credit decisions or penalize a lender for unfair lending practices. This interaction between the two can be seen as a principal–agent game where the regulator plays the role of the principal and the firm represents the agent, which we formalize in Section IV.

In our framework, there is a *training phase* in which the firm selects the algorithmic policy based on some initial data and a subsequent *deployment phase* in which it is implemented. For example, a lender uses historical repayment data to train a model in the training phase, and then leverages this model to decide who to approve for a loan in the deployment phase. This perspective on the genesis of algorithmic decisions suggests three targets for regulatory intervention (Figure 1): First, regulatory interventions could impose *ex-ante* restrictions that target the training phase and its inputs, such as which training data the firm is allowed to use, how it can be processed, and the types of models that can be trained on it. Second, the regulator could formulate *ex-interim* regulation that concerns the actual algorithmic policy after it is computed, but before it is deployed. Finally, *ex-post* regulation could target the actual decisions and outcomes that follow the implementation of the algorithmic policy.

To understand the full range of legal strategies available for regulating algorithmic systems, we distinguish between two dimensions of regulation along the temporal dimensions: *rule timing* and *scrutiny timing*. Rule timing refers to the stage in the algorithmic lifecycle that a legal rule is designed to govern. For example, rule timing distinguishes between restrictions to the inputs that can be used for training (an ex-ante rule) and requirements for a model to undergo stress testing before deployment (an ex-interim rule). Scrutiny timing, by contrast, refers to when compliance with the rule is actually assessed. These two dimensions are conceptually distinct: an ex-ante rule, such as an input restriction, could be scrutinized at different points–before training (ex-ante scrutiny, e.g., during licensing), later during model development (ex-interim scrutiny, e.g., through submitted documentation), or even after deployment (ex-post scrutiny, e.g., through post-hoc audits). Similarly, an ex-interim rule (such as a stress-testing requirement for the chosen model) may be scrutinized ex-interim (before deployment) or ex-post (only after deployment).

Critically, the separation between rule timing and scrutiny timing allows for the design of combined regimes or conditional scrutiny: a regulator may decide to evaluate compliance with an ex-ante or ex-interim rule only if an ex-post issue arises. For example, an ex-ante or ex-interim rules could be scrutinized only conditional on an ex-post rule being violated (e.g., post-hoc audits of allowed inputs in response to realized harm). Such distinctions frame the legal strategies we explore in greater detail in the taxonomy in Section III.

III. A Taxonomy of Emerging Regulatory Approaches

This section introduces a temporal taxonomy of regulatory approaches. Building on our framework from Section II, we categorize regulatory approaches by the stage of the algorithmic pipeline they target and the timing at which compliance is assessed. By making this structure explicit, we clarify how emerging regulatory proposals, as well as longstanding doctrines such as disparate impact liability, fit into this framework, and how regulators may choose between upstream prevention and downstream enforcement. As we show, algorithmic systems create opportunities for ex-interim regulation, which is limited in traditional human decision-making contexts. Moreover, scrutiny timing offers additional flexibility: even when interim scrutiny is costly or infeasible, regulators can use ex-post scrutiny of ex-interim rules, using realized outcomes to trigger more targeted assessments of whether earlier constraints were met.

The taxonomy below outlines the temporal options and illustrates how different regulatory instruments, ranging from mandates to evaluation obligations, populate this design space. While ex-ante rules can theoretically be scrutinized at any stage, practical constraints limit when other rules can be meaningfully assessed. For instance, rules governing deployment outcomes can only be evaluated after deployment has occurred. These temporal constraints create the pattern of available regulatory options shown in Table 1, where certain combinations are marked as "N/A" because they are temporally impossible.

Scrutiny timing → Rule timing ↓	**ex-ante** (before training)	**ex-interim** (during development)	**ex-post** (after deployment)
ex-ante (before training)	Rule targets ex-ante stage and is scrutinized ex-ante. Example: Input restrictions checked for licensing.	Rule targets pre-development stage but is only scrutinized at development stage. Example: Verifying compliance with input restriction through a conformity assessment before model deployment.	Rule targets pre-development stage but is only scrutinized ex-post. Example: Verifying compliance with an input restriction through post-hoc analysis after deployment harm has materialized.
ex-interim (during development)	N/A	Rule targets development stage and is scrutinized pre-deployment. Example: An explainer of model weights documented and reported pre-deployment.	Rule targets development stage but is only scrutinized at deployment. Example: Evaluation of whether inappropriately trained model parameters caused disparities at deployment.
ex-post (after deployment)	N/A	N/A	Rule targets outcomes and is scrutinized at deployment. Example: Deployment impact analysis showing true impact of model on different groups.

Table 1: Temporal taxonomy of regulatory options.

While the choice of *what* and *when* to regulate—ex-ante, ex-interim, or ex-post—determines the timing of

intervention or scrutiny, it does not dictate the *type* of regulatory tool used. These distinctions, though not the focus of our chapter, present another important consideration for regulatory design: across all points in the rule/scrutiny timeline, regulators can employ either mandates or evaluation obligations. *Mandates* are rules that prescribe specific conduct. These may take the form of clearly defined requirements, such as prohibitions on using certain inputs, or more flexible standards that still require compliance with a concrete obligation, like ensuring dataset representativeness.

Evaluation obligations, by contrast, do not typically dictate particular actions. Instead, they require firms to engage in processes such as identifying risks, documenting model development, or demonstrating that certain factors were considered in decision-making (Kaminski, 2023; Engler, 2023).[2] The remainder of this section examines how these rules can be deployed at different stages of algorithmic development and decision-making.[3]

III.1 Ex-Ante Rules

Ex-ante rules, applied prior to the development of an algorithm, cover a range of regulatory tools, including restrictions or requirements regarding future functions the firm may develop. One common type of ex-ante rule limits the inputs a firm can use when developing or training a model. For example, the prohibition on AI systems from using personal characteristics, such as social behavior or personality traits, when those characteristics were not collected for the intended use, or the prohibition on the use of emotion recognition in workplaces or educational institutions (except for medical and safety reasons) can be found in Article 5(1) of the EU AI Act (European Union, 2023). Regulation could also prohibit the development of certain models altogether, such as the EU AI Act's prohibition on social scoring, which is considered an "unacceptable risk." These restrictions all target the ex-ante stage, pre-deployment.

Below we consider how both ex-ante rules can be scrutinized at any point in the pipeline.

Ex-ante scrutiny. There are several reasons why a regulatory regime might choose to scrutinize ex-ante rules already during the pre-development stage. If the mere collection and storage of data introduce significant risks—such as when dealing with highly sensitive information like biometric data or if there is a concern about models leaking or being stolen—it may be preferable to ensure compliance before development begins rather than defer scrutiny to a later stage. Additionally, in high-risk domains that pose substantial safety concerns, regulators may seek to limit which entities are permitted to develop AI for certain purposes, requiring these entities to demonstrate that they have sufficient safeguards in place. Anderljung et al. (2023), for example, proposes requiring a governmental license for developing AI models in certain safety-critical and high-risk industries such as air travel, power generation, and manufacturing. In these examples, the regulator is more risk-averse than the model developer, creating friction that makes it undesirable to scrutinize the model at a later stage.

[2]Some legal requirements may blur the line between these categories. For example, a law might require a firm to establish a risk management system. If the law prescribes exactly how the system should be structured and what documentation must be maintained, it functions as a mandate. If it simply requires that the firm engage in risk identification and mitigation without prescribing the details, it operates more like an evaluation obligation. Article 9 of the EU AI Act (European Union, 2023), which requires a risk management system to consider foreseeable risks to "health, safety or fundamental rights," is an example of an evaluation obligation. Although the requirement applies at the ex-interim stage, it emphasizes process-oriented compliance rather than prescribing a specific outcome. Both categories of legal rules—mandates and evaluation obligations—can, in principle, correspond to any combination of rule timing and scrutiny timing represented in Table 1.

[3]Our focus here is primarily on public regulatory scrutiny, although private enforcement, such as private litigation, can also play a role in ensuring compliance. In many cases private enforcement, such as tort litigation, statutory claims or other civil recourse, operates ex-post and are often triggered by the materialization of harm. See, e.g., Sharkey (2024).

Ex-ante scrutiny often takes the form of a licensing regime, where a firm is permitted to proceed to the development stage only after meeting specific conditions. These conditions might include demonstrating compliance with input restrictions or adequately addressing potential risks through documented risk assessments.

Ex-interim scrutiny. Rather than scrutinizing an ex-ante rule during the pre-development stage, a regulator might opt to scrutinize compliance during the ex-interim development stage. For example, the EU AI Act requires deployers of high-risk AI systems to produce a report demonstrating that their systems comply with the Act's requirements (see Annex VII of European Union, 2023). In some cases, a "notified body"—a designated third party—is tasked with certifying these conformity assessments. These assessments often include ex-ante restrictions, such as prohibitions on the use of certain sensitive personal data or requirements to identify and mitigate risks at the pre-development stage. In certain circumstances, a notified body's refusal to certify the assessment may halt further development, effectively making the third-party approval an ex-interim scrutiny of certain ex-ante requirements. Similarly, Anderljung et al. (2023) suggests that certain AI models should require a government license pre-deployment, depending on the developer's ability to demonstrate compliance with certain safety standards, which could include ex-ante restrictions or requirements.

Ex-post scrutiny. Ex-post scrutiny refers to evaluating compliance with legal rules after an AI system has been deployed. Scrutinizing an ex-ante rule at this later stage may be particularly desirable when the cost of auditing and assessing compliance is high, in which case the regulator might selectively scrutinize the ex-ante rule when observing a particular outcome at deployment. For instance, a regulator might choose to consider whether a firm has complied with an input restrictions on using protected characteristics only if the deployed model results in significant racial disparities. In line with this approach, Article 79 of the EU AI Act (European Union, 2023) allows for market surveillance authorities to evaluate an already deployed AI system with respect to its compliance with the Act's obligations, covering rules that target any stage of the algorithmic life cycle, including rules that relate to pre-development. Similarly, self-assessments required for ex-ante licensing may only be scrutinized ex-post, potentially resulting in license revocation. While this approach allows regulators to focus costly enforcement efforts, it also raises challenges in cases where immediate intervention is needed to prevent ongoing or widespread harm.

III.2 Ex-Interim rules

Ex-interim rules apply to the model development stage, after training data are available but before deployment, and therefore relate to rules that restrict or constrain the algorithmic function once it is being trained. At this point, the regulator can impose rules that relate to the training data and the concrete function trained on that data. Ex-interim rules can be restrictions on the function that may depend on the training data itself. For instance, a regulator might assess whether the selected algorithmic rule is appropriate or whether a more desirable algorithmic rule could be implemented based on the information available during training, using tools such as explainers or stress-testing frameworks.[4] For example, Article 9 of the EU AI Act requires a risk management

[4]In our framework, we jointly consider obligations for both developers of AI systems and those who deploy them. However, in practice, the deployer of an AI system may not be the same as the developer. For instance, a lender (user) might use a credit scoring model created by a third party (developer). Legal frameworks such as the EU AI Act recognize this distinction and impose separate obligations on developers and deployers. For tractability, we consider these obligations jointly.

system that considers the foreseeable risks to "health, safety or fundamental rights" at deployment. Similarly, the proposed Canadian Artificial Intelligence and Data Act (AIDA) (Canada, 2024), requires that firms developing high-impact AI systems establish measures to "identify, assess, and mitigate risks of harm or biased output" before the system is deployed. The Algorithmic Accountability Act, proposed in 2023 (U.S. Senate, 2023), similarly requires concrete impact assessments before deployment (Kaminski, 2023).

Another type of ex-interim requirement puts restrictions on the training dataset—which can be understood as targeting how the data itself is processed—serve as examples of ex-interim rules. For instance, Article 10 of the EU AI Act mandates that AI systems use training, validation, and testing datasets that satisfy quality standards, such as relevance, representativeness, and absence of prohibited biases.

In practice, precisely defining the boundaries of the ex-interim and ex-post stage can be challenging, particularly when developers engage in iterative testing or smaller-scale internal evaluations that blur the line between development and deployment. This ambiguity complicates the enforcement of ex-interim rules, especially when scrutiny is intended to take place before individuals are impacted by an AI system.

As noted earlier, ex-interim rules cannot be scrutinized at the ex-ante stage because compliance with a future requirement cannot be assessed before the relevant activities have taken place. Therefore, scrutiny of ex-interim rules occurs during the ex-interim or ex-post stages.

Ex-interim scrutiny. Ex-interim scrutiny implies the evaluation of compliance with ex-interim rules during the model development stage, prior to deployment. This type of scrutiny ensures that developers adhere to requirements in real-time as the system is being developed rather than after deployment. Under the EU AI Act, providers of high-risk AI systems are required to maintain detailed documentation of their development processes, including demonstrating how training data meets specified standards. Regulators or third-party entities, such as notified bodies, may review this documentation during the development phase to verify compliance, so that these ex-interim evaluation obligations are scrutinized before deployment (ex-interim scrutiny of an ex-interim rule). Several scholars have suggested that firms should be required to demonstrate that their models are fair and non-discriminatory to receive permission to deploy the models, similar to FDA approval (Malgieri and Pasquale, 2022; Tutt, 2017).

Ex-interim scrutiny is particularly valuable for addressing risks that may be difficult or costly to mitigate after deployment. However, this approach often requires significant regulatory resources and access to the development process, which can be burdensome. As a result, regulators may prefer to allocate enforcement resources to ex-post scrutiny, where compliance is prioritized based on realized outcomes.

Ex-post scrutiny. Ex-post scrutiny refers to the evaluation of compliance with an ex-interim rule after the deployment of the AI system. This approach may be especially practical when assessing compliance with an ex-interim rule is costly or resource-intensive. In such cases, a regulator might choose to assess compliance only if harm materializes at the ex-post stage—thereby using realized harms as a trigger for selective, targeted scrutiny of ex-interim constraints.

Discrimination law often functions as an ex-post scrutiny of ex-interim decision-making. For example, while review under the disparate impact doctrine is triggered by ex-post disparities,[5] the primary consideration of

[5]EEOC v. Greyhound as an example of only caring about ex-post outcomes. The court stated that "[t]his conclusion should be as obvious as it is tautological: there can be no disparate impact unless there is a disparate impact." 635 F.2d at 191-92.

whether an entity engaged in prohibited disparate impact, for example, an employer who uses a screening tool that selects men at a higher rate than women, is whether there is a "business justification" for the policy that led to the disparity. Because this justification relates to whether the AI system was meant to predict job performance, for example, this would be scrutinizing the model development stage. Similarly, the Community Reinvestment Act triggers scrutiny of banking lending policies when financial institutions fail to adequately provide loans to low- and moderate-income areas they service. In this vein, Kleinberg et al. (2018) argue that algorithms provide an important opportunity for transparency by enabling the inspection of the training model—which they refer to as the "screener" algorithm—once decision disparities are detected, to enforce discrimination laws.

III.3 Ex-Post Rules

Ex-post rules are designed to address materialized harm rather than anticipated risks. Unlike ex-ante or ex-interim rules, which focus on preventing potential issues before deployment, ex-post rules are reactive: they assess harm that has, at least in part, already occurred. This materialized harm reflects not only the firm's algorithmic policy but also factors that are only knowable at the time of deployment and may lie beyond the firm's control. For example, a lending policy that results in a high rate of loan defaults may stem from both the lender's application policy and external factors—such as the characteristics of borrowers who apply or how economic conditions affect repayment behavior.

Rules that hold developers or deployers accountable when a deployed AI system causes physical harm or other damage—such as tort liability for autonomous vehicle accidents—are examples of ex-post rules. Particularly in strict liability regimes, where a deployer's reasonable risk-mitigation efforts do not absolve liability, the firm effectively becomes a de facto insurer against harm.

Similarly, obligations to engage in ongoing monitoring, measure outcomes, and report harms often target the ex-post stage. For example, Article 79 of the EU AI Act requires post-market monitoring in light of real-world outcomes, creating an *assessment obligation* at the ex-post stage. The proposed Canadian Artificial Intelligence and Data Act (Canada, 2024) includes provisions for monitoring and reporting incidents of harm or biased outputs from high-impact AI systems. The NIST AI Risk Management Framework (NIST, 2023), a non-binding framework in the U.S., calls for comparing system performance at deployment to the assessments made pre-deployment, reinforcing the importance of ex-post scrutiny.

Some rules combine ex-post and ex-interim elements. Ex-post rules enable regulators to focus enforcement efforts on actual harms rather than theoretical risks that may never materialize. However, liability that extends beyond the monitoring and reporting requirements—such as a determination of illegal disparate impact—may also require scrutiny of an ex-interim rule. In such cases, the ultimate finding of liability often hinges on whether the deployer or developer complied with ex-interim rules, such as demonstrating that the model's design and data usage were justified and aligned with legitimate business purposes.

Ex-post rules are less desirable in high-risk domains where regulators may prioritize preventing harm before it materializes. The trade-offs involved in harm regulation, including negligence-based and strict liability approaches, have been extensively examined in the law and economics literature (e.g. Polinsky and Shavell, 2000).

IV. A Stylized Model of Algorithm Deployment and Regulation

In this section, we formally model the regulation of algorithms within the framework from Section II as a game between the firm that deploys an algorithm and the regulator that oversees its use. The regulator can impose rules that constrain the algorithmic policies selected by the firm, with the type of restriction depending on the stage in the decision-making pipeline. Rules that target the ex-ante stage are limited to simple restrictions that do not directly rely on training data, such as restrictions on inputs or the types of models that can be trained. Rules that target the ex-interim stage focus on restricting aspects of algorithmic policies that emerge after the firm accesses the training data. Finally, rules targeting the ex-post stage impose restrictions based on the observed impact of the algorithmic policies after deployment. Additionally, the regulator must decide when to scrutinize compliance with a rule, which can also impact the cost of regulatory oversight. A key example from Section II that this model aims to capture is whether to evaluate compliance with ex-interim restrictions prior to deployment or to defer scrutiny until after deployment.

Following the classification in Section II, within our model we consider regulatory regimes that vary in their rule and scrutiny timing. We argue that the regulator's choice of the timing of the rule and its scrutiny generally depends on the relative costs of each approach, and whether the approach can sufficiently address the regulator's objectives. As an illustration, we provide a simple example to demonstrate the trade-offs between different regulatory options, highlighting the opportunity for regulation to target algorithmic decisions even before they have been deployed and possibly scrutinize compliance before harm can materialize at deployment.

IV.1 Setup

We formally model the setting from Section II as a stylized game between a firm (agent) deploying an algorithm and a regulator (principal) overseeing its use. The firm chooses an algorithmic policy f from some set $\mathcal{F}$. We think of this algorithmic policy as a mapping $f : X \to \mathcal{A}$ from features $x \in X$ to a decision $f(x) \in \mathcal{A}$. For example, the firm could be a lender who decides whether to give credit ($a \in \{\text{extend credit}, \text{deny credit}\} = \mathcal{A}$) to a borrower with financial history x. The regulator puts constraints or imposes penalties on the algorithmic decisions f. For example, a financial regulator may restrict the variables that are permissible for use in credit decisions or penalize a lender for unfair lending practices.

We assume that the firm and the regulator may have different preferences over algorithmic decisions. The firm wants to achieve high profit $\Pi_\theta(f)$, such as the net return to lending. The regulator aims to maximize utility $U_\theta(f)$, which may differ from the firm's goal, e.g. by additional fairness or risk considerations. Both objectives could depend on the state of the world $\theta \in \Theta$ when the algorithmic policy f is deployed. For example, this state of the world may determine the joint distribution of some outcome of interest y (such as repayment) and the covariates x (such as the financial histories of actual loan applicants).

We assume that there is a training phase in which the algorithmic policy is chosen and a deployment phase in which it is applied. In the *training phase*, the firm chooses an algorithmic policy $\hat{f}_\tau \in \mathcal{F}$ based on a training signal $\tau \in T$. This signal represents the training data that is used by the firm when deciding between functions $f \in \mathcal{F}$. In the *deployment phase*, the state of the world $\theta \in \Theta$ is realized, leading to lender profit $\Pi_\theta(\hat{f}_\tau)$ and regulator utility $U_\theta(\hat{f}_\tau)$. We model the idea that this training signal is informative about the deployment state by assuming that $(\tau, \theta) \in T \times \Theta$ comes from a joint distribution P. For example, the training data may be a

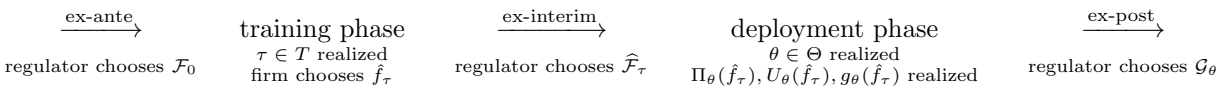

Figure 2: Rule timing of regulatory interventions.

sample from the same distribution as the deployment. We assume that both the regulator and the firm know the joint distribution P and first learn τ and later θ.

Without any regulation, a firm that maximizes expected profit would choose an algorithmic policy $\hat{f}_\tau$ during the training phase that maximizes $\mathrm{E}[\Pi_\theta(f)|\tau]$ over all available policies $f \in \mathcal{F}$, leading to realized profit $\Pi_\theta(\hat{f}_\tau)$ to the firm and utility $U_\theta(\hat{f}_\tau)$ to the regulator. However, the firm's first-best choice may have undesirable properties from the regulator's perspective, such as when a lender's credit-scoring policy excludes part of the target population or leads to excess systemic risk. The regulator, therefore, may want to restrict the firm's choice of algorithmic policy.

IV.2 Target and Timing of Regulation

Following the taxonomy in Section II, regulatory interventions in our game can target and happen at three points in time: before the training phase (ex-ante); between the training and deployment phases (ex-interim); and after deployment (ex-post). *Ex-ante*, the regulator can decide on general restrictions on how the firm is allowed to process training data that does not depend on the particular training data that is later realized. For example, the regulator may generally rule out that policies use some prohibited features. Alternatively, a rule targeting the *ex-interim* stage, after the training sample τ is available and the firm chooses its algorithmic policy, can determine whether a particular algorithmic rule $\hat{f}_\tau$ is permissible for use in the specific context. For example, a regulator concerned with financial stability may, based on the results of a stress test of a lending rule and its performance in an adverse scenario even before it is deployed, determine whether the use of a lending rule is permissible. Finally, *ex-post* rules relate to the stage at which the deployment state θ is realized and the algorithmic rule is deployed. A this stage the regulator can observe outcomes and sanction the firm based on them. For example, a banking regulator concerned with fairness may check for disparities in realized lending decisions and decide, based on those disparities, whether to penalize the lender.

In all three cases, we assume that regulatory rules define a set of permissible algorithmic policies $\hat{f}_\tau$ that the lender can choose without facing sanctions. However, the nature of these restrictions differs across different types of regulation.

Ex-ante rules make simple general restrictions about the way that algorithmic rules can be chosen in the first place, such as restrictions on how training data is allowed to be used or which algorithmic rules can be searched over in the first place. For simplicity, we focus on one specific way of capturing such restrictions: we assume that ex-ante rules simply limit the permissible algorithmic policies to some fixed set. That is, these rules take the form $\hat{f}_\tau \in \mathcal{F}_0$ for some ex-ante chosen $\mathcal{F}_0 \subseteq \mathcal{F}$. This captures, for example, that specific characteristics cannot be used by the final policy, like an input restriction, or that some model classes are not allowed to be considered.[6]

[6]There are alternative options for capturing the idea of rigid ex-ante rules from our taxonomy in Section III. For example, we could restrict the way that training data is allowed to be processed, and limit the firm to only use some coarser information CleanedData(τ), where the regulator chooses the mapping CleanedData. Or we could think more generally of putting some coarse,

For *ex-interim rules*, restrictions are directly on the function $\hat{f}_\tau$, and may depend on the training data itself. That is, these rules take the form $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau \subseteq \mathcal{F}$. In this case, the regulator checks whether the selected algorithmic rule is appropriate, potentially using the training data to make this determination. For example, a regulator may check whether a better algorithmic rule *is* available (when scrutinized ex-interim) or *would have been* available (when scrutinized ex-post) based on the information available during training.

For *ex-post rules*, tests take the form whether $g_\theta(\hat{f}_\tau) \in \mathcal{G}_\theta$, where $g_\theta(f)$ are some realized outcomes of deploying a function $f \in \mathcal{F}$ in state θ. That is, ex-post rules only depend on the implications of the algorithmic rule in the realized state θ. Here, the function $g_\cdot(\cdot)$ is fixed and predefined, while the choice of the set $\mathcal{G}_\theta$ expresses the regulator's actual restriction in the realized world. When there are disparity concerns, then $g_\theta(f)$ may capture the realized differences in lending rates across protected groups, and $\mathcal{G}_\theta$ the range of differences the regulator is willing to accept ex-post.

Figure 2 situates each of these three rules, represented by the choices $\mathcal{F}_0, \widehat{\mathcal{F}}_\tau, \mathcal{G}_\theta$, within the overall timeline.

We assume that determining the legal rule and compliance with it can be costly to the regulator, and that ensuring $\hat{f}_\tau \in \mathcal{F}_0$ costs $c_{\text{ex-ante}}$, that testing $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$ comes at a cost $c_{\text{ex-interim}}$, and that checking $g_\theta(\hat{f}_\tau) \in \mathcal{G}_\theta$ comes at a cost $c_{\text{ex-post}}$.[7]

Following our general framework, we distinguish between the *rule timing*—that is, which information the specific rule is based on—and the *scrutiny timing*—that is, when compliance with the rule is assessed. We assume here that the cost of testing some rule only depends explicitly on the timing of the rule, but not on the timing of the scrutiny. That is, whether an ex-interim rule is scrutinized ex-interim or ex-post comes at the same cost. However, ex-post scrutiny of an ex-interim rule may still be cheaper if it is performed only when bad outcomes materialize, thus reducing the expected cost.

We assume that ex-interim scrutiny implies that the regulator forces the firm to choose a permissible function $\hat{f}_\tau$ that complies with ex-ante and ex-interim rules, whereas for ex-post scrutiny the regulator decides whether to allow the firm to collect profits $\Pi_\theta(\hat{f}_\tau)$ or to punish the firm by reducing profits to $\underline{\Pi}$. The timing of the scrutiny, therefore, matters for the overall cost to the regulator. Ex-post scrutiny of ex-interim rules—such as when a regulator uses materialized harm for targeting auditing efforts of ex-interim compliance—may involve the cost $c_{\text{ex-post}}$ of testing ex-post as well as the cost $c_{\text{ex-interim}}$ of ex-interim testing, for firms that are selected for auditing. Throughout, we consider the case where the regulator commits to a deterministic regulatory policy (in the form of ex-ante rules $\mathcal{F}_0$, ex-interim rules $\tau \mapsto \widehat{\mathcal{F}}_\tau$, ex-post rules $\theta \mapsto \mathcal{G}_\theta$, and the decision when to scrutinize each).[8]

IV.3 Regulatory Regimes

We now make different regulatory options concrete by listing five specific regulatory regimes that mirror our taxonomy in Section III. In the example below, we consider the case where ex-ante rules can be scrutinized for free, $c_{\text{ex-ante}} = 0$, which expresses the idea that simple restrictions on inputs or functional form may be easy to verify. Typically, we would assume that ex-post rules are cheap to enforce while checking ex-interim rules may be

easy-to-implement restrictions on the mapping $\tau \mapsto \hat{f}_\tau$.

[7] A more general formulation would make the costs depend on the specific restrictions; for example, $c_{\text{ex-ante}}$ may depend on the specific restriction $\mathcal{F}_0$, as some restrictions may be cheaper to verify than others.

[8] Natural extensions include cases where the regulator imposes a fine in the ex-post and possibly also in the ex-interim stage, cases with audits that only happen with a given probability, cases with limited commitment, and repeated interactions where an infraction in a first interaction leads to additional regulatory obligations in the next.

more costly ($c_{\text{ex-interim}} > c_{\text{ex-post}} \geq 0 = c_{\text{ex-ante}}$) since they require inspecting the complex algorithm (and possibly the training data) beyond the realized consequences of algorithm deployment. The five regulatory regimes we consider are:

1. No regulatory constraints: In the laissez-faire regime, there are no regulatory constraints, and the firm is free to choose any $\hat{f}_\tau \in \mathcal{F}$ after observing the training signal τ.

2. Ex-ante restriction: If the regulator imposes ex-ante rules $\mathcal{F}_0$, the firm is forced to choose $\hat{f}_\tau \in \mathcal{F}_0$, but otherwise unrestricted. (Here, since we assume that ex-ante rules can be checked without cost to the regulator, we assume that they can also be perfectly enforced; in particular, we do not consider different scrutiny timing options.)

3. Ex-post rule: If the regulator imposes ex-post rules, then the regulator does not put any ex-ante or ex-interim restrictions. Instead, the regulator can pay $c_{\text{ex-post}}$ to choose $\mathcal{G}_\theta$. In this case, the firm obtains profit $\Pi_\theta(\hat{f}_\tau)$ if $g_\theta(\hat{f}_\tau) \in \mathcal{G}_\theta$ and $\underline{\Pi}$ otherwise. The regulator obtains utility $U_\theta(\hat{f}_\tau) - c_{\text{ex-post}}$ since the algorithmic policy is already deployed.

4. Ex-interim audit: If the regulator performs an ex-interim audit, the regulator pays $c_{\text{ex-interim}}$ and chooses $\widehat{\mathcal{F}}_\tau$, and the firm chooses $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$. This case corresponds to ex-interim scrutiny of an ex-interim rule.

5. Ex-interim rule scrutinized by ex-post audit: If the regulator imposes ex-interim rules based on an ex-post audit, the regulator first tests based on an ex-post rule $g_\theta(\hat{f}_\tau) \in \mathcal{G}_\theta$, at cost $c_{\text{ex-post}}$, whether to perform an in-depth audit. If the rule is violated, the regulator pays an additional $c_{\text{ex-interim}}$ to assess whether $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$. In this case, the firm obtains profit $\Pi_\theta(\hat{f}_\tau)$ if $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$ and $\underline{\Pi}$ otherwise. The regulator obtains utility $U_\theta(\hat{f}_\tau)$ net of the costs of the audits. Writing δ for the probability that the ex-post rule is violated ($g_\theta(\hat{f}_\tau) \notin \mathcal{G}_\theta$), this means that the expected cost of the regulator is $c_{\text{ex-post}} + \delta c_{\text{ex-interim}}$. This option corresponds to ex-post scrutiny based on ex-interim and ex-post rules in the parlance of Section II.

While these regulatory options could be combined in principle, we consider them separately to simplify our analysis. Throughout, we assume that the regulator and the firm maximize expected utility and profit, respectively.

IV.4 Regulation in an Illustrative Example

In order to illustrate insights from our general model, we now provide a simple instance. Specifically, we consider a lender who selects between three decision rules for allocating credit, namely a risky rule that expands credit to many borrowers, a conservative rule that provides credit only to very safe borrowers, and an imprecise rule that adds unnecessary noise, $f \in \mathcal{F} = \{\text{risky}, \text{conservative}, \text{imprecise}\}$. We assume that the deployment state can either be high or low, $\theta \in \{\text{high}, \text{low}\}$, which could represent, for example, a period of financial stability (high) or instability (low). The immediately observed outcome can be one of $g_\theta(f) \in \{\text{great}, \text{good}, \text{bad}\}$.

The lender generally prefers the risky to the conservative policy. The financial regulator prefers expanding credit access in the high state, but is concerned about harm to marginal borrowers who cannot repay their loans in the low state. So the regulator prefers the risky rule to be used in the high state and the conservative rule

	Lender profit $\Pi_\theta(f)$		Regulator utility $U_\theta(f)$		Realized outcome $g_\theta(f)$	
	$\theta =$ high	$\theta =$ low	$\theta =$ high	$\theta =$ low	$\theta =$ high	$\theta =$ low
$f =$ risky	6	2	5	-1	great	bad
$f =$ conservative	4	1	2	2	good	good
$f =$ imprecise	2	0	1	0	bad	bad

Table 2: Lender and regulator payoffs in the example.

in the low state. Neither the regulator nor the lender prefers the imprecise rule. Concrete profits and regulator utility that represent these rank-orderings are provided in Table 2.

In the training phase, we assume that the lender receives a noisy signal $\tau \in T = \{\text{likely high}, \text{certainly low}\}$ about the future deployment state, with $\mathrm{P}(\tau = \text{certainly low}) = 1/4$. (These signals represent training data that are indicative of the repayment probabilities of borrowers in the deployment phase.) If the training signal is $\tau =$ certainly low, then the deployment state will be low, $\mathrm{P}(\theta = \text{low}|\tau = \text{certainly low}) = 100\%$. If it is $\tau =$ likely high, then there is a 2/3 chance of the high state occuring, so $\mathrm{P}(\theta = \text{high}|\tau = \text{likely high}) = 2/3$. This distribution is also represented in Figure 3.

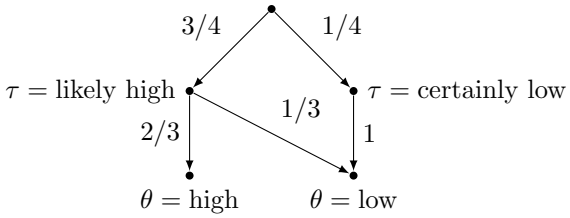

Figure 3: Joint distribution of training and deployment signals in the example.

The lender observes the training signal τ and chooses an algorithmic policy $\hat{f}_\tau$ to maximize expected profit, subject to potential regulatory constraints. If the training signal is $\tau =$ likely high, then regulator and lender agree that the risky algorithmic policy is optimal. But if $\tau =$ certainly low, then the regulator prefers the conservative policy, while the lender still prefers the risky policy.

We now consider the different regulatory regimes from Section IV.3 in this example. Throughout, we assume that the regulator aims to maximize expected utility, that the regulator can punish the lender by setting profit to $\underline{\Pi} = -6$ (e.g. by excluding it from making profits in the future), that the cost for ex-interim tests is $c_{\text{ex-interim}} = c$ with $0 < c < 3$, and that ex-ante and ex-post tests are not costly ($c_{\text{ex-ante}} = 0 = c_{\text{ex-interim}}$). The different regulatory regimes then play out as follows:

1. No regulatory constraints: In the laissez-faire regime without regulatory constraints, the lender always chooses the risky algorithm, leading to expected lender profit of 4 and regulator utility of 2.

2. Ex-ante restriction: If the regulator can only impose ex-ante rules of the form $\hat{f}_\tau \in \mathcal{F}_0$, then the regulator could force the lender to avoid the risky rule (and pick the safe rule instead), $\mathcal{F}_0 = \{\text{conservative}\}$. With our parameters, this would still lead to safe regulator utility of 2 (but reduce expected lender profit to 2.5).

3. Ex-post rule: If the regulator imposes an ex-post rule, they can punish the lender as a function of the outcome in the deployment state. In the example, a plausible option would be to punish the lender (by

imposing profit $\underline{\Pi} = -6$) whenever the bad outcome materializes ($g_\theta(\hat{f}_\tau) = \text{bad}$). In this case, the lender would not choose the risky option in either the high or the low training state, since even in the high training state the expected profit (net of regulatory sanctions) from deploying the risky algorithmic rule is now 2, relative to an expected profit of 3 for deploying the conservative policy. Hence, regulator profit would still be 2, but lender profit would be reduced to 2.5.

Of course, the effectiveness of these ex-post audits depends on the severity of punishment. If, for example, the regulator could impose a lighter punishment of, say $\underline{\Pi} = 0$, then this policy could effectively rule out the choice of the risky algorithmic option in the low training state only. However, such a policy would still be limited in a different instance of our model: if there is a possibility of the high state occurring even after a low training state, then limited liability by the lender may make ex-post sanctions ineffective since they may not be enough to force the choice of the conservative algorithmic policy even in the low training state.

4. Ex-interim audit: If the regulator intervenes at the interim stage, they can pay a fee of c to impose restrictions on the choice of the lender based on the training signal. Here, this would imply a restriction to the regulator's preferred choice,

$$\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau = \begin{cases} \{\text{risky}\}, & \tau = \text{likely high}, \\ \{\text{conservative}\}, & \tau = \text{certainly low}. \end{cases}$$

In this case, the overall expected regulator utility, net of cost, is $2.75 - c$, which is an improvement over the other options as long as the cost is small enough ($c < .75$). The expected lender profit in this case is 3.75.

5. Ex-interim rule scrutinized by ex-post audit: The regulator could also intervene in the ex-post stage, but still use ex-interim information. In this case, a natural policy would be to only scrutinize the training phase if the risky algorithmic policy was deployed in the low state. In this case, if the regulator threatens to pay a cost c to check whether the lender chose $\hat{f}_\tau = \text{risky}$ for $\tau = \text{certainly low}$, but only in the case that $\theta = \text{low}$ and $\hat{f}_\tau = \text{risky}$ are realized, then it would push the lender to implement the preferred choice

$$\hat{f}_\tau = \begin{cases} \text{risky}, & \tau = \text{likely high}, \\ \text{conservative}, & \tau = \text{certainly low}. \end{cases}.$$

However, the cost is now lowered to $c/4$, since the ex-interim scrutiny only happens in the case of a low deployment state following the optimistic training signal, for an expected regulator utility (net of cost) of $2.75 - c/4$. Expected lender profit is unchanged at 3.75.

This specific example highlights the potential of targeted ex-interim rules for algorithmic regulation, even in cases where they are more costly to enforce than ex-ante and ex-post rules. In addition, disentangling rule and scrutiny timing can yield regulatory solutions with lower cost.

IV.5 Frictions from Algorithmic Complexity

Our approach to ex-interim regulation of algorithms is built around tests of the form $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$ that check whether the algorithmic decision function belongs to some permissible set. In practice, the algorithmic policies chosen by the firm may be very complex. For example, a lender may utilize deep neural networks or other machine learning

methods to determine who should receive credit. In such cases, the regulator may not be able to parse all the nuances of the firm's choice $\hat{f}_\tau$, or the firm may be limited in its ability to share all the algorithm's details for privacy or intellectual property reasons. Instead, the regulator may have to rely on simplified (typically low-dimensional) descriptions of algorithmic decisions. In practice, these could take the form of variable-importance measures, simpler proxy models, or evaluations at a limited number of data-points. In our model, we can capture the restriction that the regulator may not be able to fully capture complex AI algorithms by restrictions on SimpleDescirption($\hat{f}_\theta$) rather than on $\hat{f}_\theta$. Concrete examples from our previous work include audits based on simple explainers (Blattner, Nelson, and Spiess, 2024) and "discrimination stress testing" based on assessing an algorithmic policy by evaluating it at specific test points (Gillis and Spiess, 2019).

In the stylized example from Section IV.4 above, such frictions could imply that the regulator cannot distinguish between all algorithmic policies in $\mathcal{F}$, and instead only observes whether a given choice falls within a coarse partition. *Which* functions the regulator can distinguish between then depends on the technology used to explain or describe the algorithm and its output (Blattner, Nelson, and Spiess, 2024). For example, we could consider different coarse ways of describing algorithms, each of which partitioning the space $\mathcal{F} = \{\text{risky}, \text{conservative}, \text{imprecise}\}$ into two non-trivial parts. A first way focuses on the overall behavior of the algorithmic policy. If the risky and conservative options are overall more similar to each other than to the imprecise policy, then such a coarsening may make it impossible for the regulator to distinguish between $f \in \{\text{risky}, \text{conservative}\}$, which would make effective regulation impossible. Instead, an effective way of summarizing complex functions has to be mindful of the source of preference misalignment. Here, this would mean that a coarsening into, say, {risky} and {conservative, imprecise} would do the job, since it preserves all the information relevant to the misalignment between lender and regulator.

V. General Implications for Regulating Algorithms

The prior sections introduced a framework of a regulator leveraging ex-ante, ex-interim, and ex-post tools to regulate a firm that deploys an algorithmic policy. In this section, we highlight three key implications within the context of the formal model in Section IV, and discuss them in connection with other related work.

V.1 The Value of Ex-Interim Regulation

First, we observe that the ex-interim stage—something we argue is unique to the *algorithmic* context—has value from a regulatory perspective. To see this, contrast the option of ex-interim intervention with several possible ex-post regulatory regimes. One such ex-post regime is to penalize the firm for undesirable outcomes. While this helps align preferences, it risks penalizing the firm for deployment-state realizations beyond its control, such as external market conditions. This undermines the primary goal of many legal rules, which is to target specific unwanted conduct rather than impose strict liability for all adverse outcomes. The regulator then faces a trade-off, as in a classic delegation problem, between a penalty being either so mild that it leads to excessive realizations of bad outcomes from the regulator's perspective, or so strict that it leads to overly conservative behavior by a risk-averse firm. Ex-interim regulation, in contrast, allows the regulator, in principle, to target precisely the aspects of the firm's conduct that are under the firm's control.

In our framework, ex-interim rules can even lead to first-best outcomes for the regulator. As an extreme

case, if there is no cost to ex-interim tests ($c_{\text{ex-interim}} = 0$ in the model from Section IV), then ex-interim audits generally dominate ex-ante and ex-post rules. First, ex-interim tests of the form $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$ are strict generalizations of ex-ante tests of the form $\hat{f}_\tau \in \mathcal{F}_0$. Second, relative to ex-post tests, ex-interim rules are more precise because they directly scrutinize the use of training data, rather than relying on the (possibly noisy) realization of the deployment state over which the firm has no control. Third, ex-interim rules can be applied before deployment, which avoids bad outcomes before they can happen. In the extreme case when there are also no frictions in imposing restrictions on algorithmic policies, the regulator can ensure that their first-best choice is deployed by performing an ex-interim audit with $\widehat{\mathcal{F}}_\tau = \{\hat{f}^*_\tau\}$ with $\hat{f}^*_\tau \in \arg\max_{f \in \mathcal{F}} \mathrm{E}[U_\theta(f)|\tau]$. In this world, the regulator effectively chooses the algorithmic policy.

The opportunity of the ex-interim stage for regulation is also pointed out by Kleinberg et al. (2018), which argues that ex-interim rules in algorithmic settings offer a distinctive advantage by leveraging the transparency and specificity of algorithmic processes, enabling clear attribution of disparities to particular modeling choices. When the regulator either cannot observe the training state or cannot fully understand the algorithm, optimal regulation is more subtle. Blattner, Nelson, and Spiess (2024) analyze a setting similar to that in Section IV above and characterize when it is optimal for a regulator to use ex-interim regulation in the form of an *explainer* that, by projecting the firm's algorithmic policy into a lower-dimensional representation that shows how it behaves in a few (carefully chosen) dimensions, can constrain the firm's behavior in ways targeted at the principal–agent preference misalignment; in general, this leads to second-best outcomes unless misalignment is very severe and high-dimensional. Gillis and Spiess (2019) develop a related ex-interim regulatory approach that involves deploying the algorithmic policy on test data—an approach they refer to as "stress testing" an algorithm. Inspired by bank stress testing, under this approach, a model is tested under a hypothetical materialization of the deployment state θ that is unknown to the developer during model training. In contrast with Blattner, Nelson, and Spiess (2024), whose explainer tool can be thought of as examining ex-interim how an algorithmic policy behaves on several columns (variables) of data, the Gillis and Spiess (2019) approach can be thought of as examining algorithm behavior on several rows (individual observations) of data. This has the advantage of potentially being easier to implement than an explainer and may also be a well-suited tool when misalignment is best summarized by different preferences over a few (high-dimensional) examples rather than by a few data features.

Another consideration is that the form of optimal regulation depends on the relationship between the training and deployment state, as well as the effectiveness of ex-post punishments. In settings where training and deployment data are very similar, scrutiny based on ex-interim rules may be unnecessary and scrutiny based on ex-post outcomes may be sufficient, provided that effective ex-post punishment leads to high compliance with the ex-post rule. On the other hand, if there is substantive additional uncertainty about the deployment state, then optimal regulation is likely to include ex-interim rules. If there is limited room for ex-post punishment, then such rules should also be enforced at the interim stage.

More broadly, the distinction between ex-interim and ex-post regulation parallels the distinction between certification regimes and enforcement regimes. Consider how the US regulates car safety, which at a high level can be viewed as a combination of (1) supervising car manufacturers' car development process and certifying a car model as meeting safety standards at the pre-production stage, and (2) periodically investigating cars' safety performance in the "deployment" state of being driven, either through annual safety inspections (as mandated in some US states) or through NHTSA-ordered recalls (at the federal level). (*Ex-ante* regulation of cars, meanwhile,

might take the form of restrictions on which types of cars can even be designed, for example, maximum allowable axle length.) Algorithmic decision-making enables such ex-interim regulation to be used for a broader set of economic activity.

V.2 The Optimal Form of Ex-Post Scrutiny

A second implication of our framework is that, even if regulation is more feasible or cost-effective to apply ex-post—for example, because of the cost $c_{\text{ex-interim}}$ of formulating rules that involve the training data in the formalization of Section IV—any ex-post regulatory tools will optimally also include ex-interim information. This follows from a logic similar to that in the preceding subsection: effective regulation should optimally target the actual conduct, rather than only possibly noisy consequences. Section III points to some real-world examples of such interventions, and Section IV provides a theoretical illustration.

There are cases where ex-post scrutiny based on ex-post rules alone can be effective or desirable. For example, if the regulator in an extension of our model would be able to impose arbitrary penalties, then they could align preferences based on realized outcomes alone (e.g. by imposing a penalty $\Pi_\theta(\hat{f}_\tau) - U_\theta(\hat{f}_\tau)$). However, if the firm has limited liability, transfers are limited otherwise, or it is infeasible to observe or estimate the necessary utility differences even ex-post, then effective regulation without ex-interim scrutiny may become inefficient.

V.3 Cases for Ex-Ante Regulation

When might ex-ante regulation still be valuable in the algorithmic setting? A third important takeaway from our framework is an understanding of cases in which ex-ante regulation may be valuable. To recap, in our framework, ex-ante restrictions refer to limitations on algorithms that do not rely on any training or deployment data. This ex-ante regulation can take several forms. In our model, they capture the ex-ante exclusion of specific inputs and models before training starts. In a broader sense, we could think of simple ex-ante restrictions as easily enforceable high-level restrictions to the data-processing and training process.

The two leading cases in which ex-ante regulation makes sense within our framework are when the cost of implementing ex-post and especially ex-interim rules is very high (such as when accessing training and deployment data is complicated), and when misalignment can be easily captured by preferences over which inputs to use (such as when the regulator wants to enforce that only certain variables are used for pricing by a firm). But when ex-interim and ex-post rules are easily enforced, then they subsume any ex-ante rules. However, the trade-offs between ex-ante and downstream rules can become more complex when additional frictions are involved.

In a related model that introduces frictions through algorithm complexity, Blattner, Nelson, and Spiess (2024) considers when it might be optimal to ex-ante restrict an algorithmic agent to use only simple models. Their analysis points to several such conditions. First, if the loss from restricting to simple models is small, or if misalignment between regulator and lender is particularly severe (in a sense Blattner, Nelson, and Spiess, 2024 formalizes), then ex-ante restrictions to simple models can be beneficial. Second, if the regulator cannot observe the training state and if their regulator's prior is relatively uninformative about the training state, ex-interim tools may have less use. Concerns about artificial intelligence presenting an existential risk for society (e.g., Jones, 2023) often reflect some combination of these conditions: an uninformative prior about the training of AI tools, or a perception of severe misalignment, or a belief in the loss from using simple models being modest.

A specific case of ex-ante regulation involves restricting which inputs an algorithm can access. Typically, such input restrictions are considered in the context of anti-discrimination regulation for protected characteristics, such as race or gender. While input-based regulation is often considered fraught in the era of "big data" (Kleinberg et al., 2018; Gillis, 2021), given how high-dimensional combinations of permitted inputs can be used to proxy for a forbidden input (Mullainathan and Spiess, 2017), Liang et al. (2021) makes the case for input-based restrictions to achieve certain fairness goals. Relative to Blattner, Nelson, and Spiess (2024), the setup in Liang et al. (2021) highlights several conditions that contribute to ex-ante restrictions being valuable: the principal has limited uncertainty about the training or deployment state, so that the form of optimal regulation is relatively knowable from the ex-ante perspective; and the principal and agent have particularly strong misalignment (in the form of strong fairness preferences of the regulator).[9]

V.4 Other Considerations

Our framework necessarily excludes some features that may also be relevant for the context-specific regulation of algorithms. One crucial consideration not included above is the demand for privacy and other normative considerations in algorithmic regulation. Ex-ante restrictions on specific model inputs, for example, may be justified either by privacy concerns or by concerns about whether including these features *per se* in a model would be ethically undesirable (e.g., Goldfarb and Tucker, 2012; Acquisti, Taylor, and Wagman, 2016; Kiviat, 2019; Goldfarb and Tucker, 2024).

In addition, our model focuses on the regulator scrutinizing the algorithmic policy actually chosen by the firm. Some regulatory approaches focus instead on the *procedure* that generated these algorithmic decisions from training data, beyond pre-processing of the data. This procedure—that is, the actual algorithm—is represented in our model as the mapping from training data to algorithmic policy, $\tau \mapsto \hat{f}_\tau$. The regulator in our model scrutinizes the algorithm's output $\hat{f}_\tau$ for the *realized* training data τ, rather than the full mapping from *any* training data τ to the algorithmic policy $\hat{f}_\tau$. In our model, we focus on the chosen policy since only this realization ultimately gets deployed and enters the regulator's utility. In addition, in practice, training algorithms may involve manual steps that are hard to capture fully. Nevertheless, there could be cases where scrutiny already happens *before* training data is available or in which communicating the training data may be infeasible. In those cases, scrutinizing how training data is generally processed—that is, analyzing the full mapping $\tau \mapsto \hat{f}_\tau$—may be part of effective regulation as a variant of our ex-interim approach.

Another relevant question we have not addressed is how the regulator learns about the markets they are regulating, which may be particularly relevant in some emerging algorithmic settings where regulatory precedent is scant. The regulator's data collection may also interact with competing firms' incentives to differentially disclose data to the regulator.[10] We also do not consider the question of how frequently to audit a firm, or how to target these audits (e.g., the resource-constrained regulator's problem of whether to audit a small firm with probable rule violations, or a larger firm with less likely, but potentially more widespread, rule violations).

Finally, we only consider the decision of the firm on how to turn training data into algorithmic policies, but not their decision on which training data to acquire in the first place. Questions of fairness can be decided as

[9] An interesting intermediate case is whether input restrictions can productively be combined with the Blattner, Nelson, and Spiess (2024) explainer tool. For example, if the explainer that a regulator has access to is too low-dimensional to sufficiently capture the misalignment with the firm, but the regulator additionally has access to ex-ante restrictions on model inputs, then combining an ex-interim explainer that targets some model features with an ex-ante restriction on using other model features may be optimal.

[10] For an interesting related analysis, see work in Callander and Li (2024).

much by which data is used as by how an algorithm is trained (Blattner and Nelson, 2024; Caro, Gillis, and Nelson, 2025), and obtaining high-quality data that represent the deployment distribution well may be a costly investment that firm and regulator have different preferences over.

VI. An Agenda on Building Tools for AI Regulation

The unique opportunity for novel approaches to regulation in algorithmic settings also suggests a broad agenda for future research. We close with a few highlights from what we see as a wide-open area.

On the one hand, this chapter argues that there are already clear opportunities for the effective regulation of algorithms. Among those is that regulation in an algorithmic setting can intervene at an *ex-interim* stage where a firm has designed an algorithmic policy but has not yet deployed it. The framework developed here, building on some of our other work (e.g., Blattner, Nelson, and Spiess, 2024; Gillis and Spiess, 2019), helps articulate why this ex-interim stage has value. Our framework also shows the broad—but not universal—conditions under which a regulator would prefer ex-interim regulation over more traditional ex-ante or ex-post interventions. Drawing on insights from the landscape of emerging AI regulation in practice, we also show how the option to delay the *scrutiny* of these ex-interim processes until a later ex-post stage can have added benefit beyond an ex-interim approach alone.

On the other hand, we see a long list of unanswered questions, including some related to how our conclusions above extend (or do not) to settings with other features than those considered here. Among other open issues discussed in Section V, we see it as particularly exciting to understand how the demand for privacy or other ethical considerations interact with algorithmic regulation, how the algorithmic development process itself might be regulated distinctly from the trained algorithmic policy, and how competitive forces and dynamic incentives among firms might shape the information available to regulators at the ex-ante stage when deciding their approach to regulation, and to firms at the training stage when they compute algorithmic policies.

Taking a step back from the economics of regulation, we also see it as important to ask what market solutions might be available for some of the regulatory challenges studied here. For example, could a market emerge for ex-interim certification, and under what conditions? What might be the advantages and disadvantages of having a private firm provide such certification services in lieu of a regulator? And—*quis custodiet ipsos custodes*—would this market for certification itself benefit from regulation?

References

Acemoglu, Daron (2021). Harms of ai. Technical report, National Bureau of Economic Research. (Cited on page 4.)

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman (2016). The economics of privacy. *Journal of economic Literature*, 54(2):442–492. (Cited on pages 4 and 20.)

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb (2023). Do we want less automation? *Science*, 381(6654):155–158. (Cited on page 4.)

Ali, S Nageeb, Greg Lewis, and Shoshana Vasserman (2023). Voluntary disclosure and personalized pricing. *Review of Economic Studies*, 90(2):538–571. (Cited on page 4.)

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. (2023). Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*. (Cited on pages 4, 7, and 8.)

Bar-Gill, Oren, Cass R Sunstein, and Inbal Talgam-Cohen (2023). Algorithmic harm in consumer markets. *Journal of Legal Analysis*, 15(1):1–47. (Cited on page 4.)

Bergemann, Dirk, Benjamin Brooks, and Stephen Morris (2015). The limits of price discrimination. *American Economic Review*, 105(3):921–957. (Cited on page 4.)

Blattner, Laura and Scott Nelson (2024). How Costly is Noise? Data and Disparities in Consumer Credit. (Cited on page 21.)

Blattner, Laura, Scott Nelson, and Jann Spiess (2024). Unpacking the black box: Regulating algorithmic decisions. *arXiv preprint arXiv:2110.03443*. (Cited on pages 4, 17, 18, 19, 20, and 21.)

Callander, Steven and Hongyi Li (2024). Regulating an innovative industry. Technical report, National Bureau of Economic Research. (Cited on pages 4 and 20.)

Canada (2024). Artificial intelligence and data act (aida). Proposed Legislation. (Cited on pages 9 and 10.)

Caro, Spencer, Talia Gillis, and Scott Nelson (2025). Modernizing fair lending. *Journal of Legal Analysis*, Forthcoming. (Cited on page 21.)

Comunale, Mariarosaria and Andrea Manera (2024). The economic impacts and the regulation of ai: A review of the academic literature and policy actions. (Cited on page 4.)

Cowgill, Bo and Catherine E Tucker (2020). Algorithmic fairness and economics. *Columbia Business School Research Paper*. (Cited on page 4.)

Dubé, Jean-Pierre and Sanjog Misra (2023). Personalized pricing and consumer welfare. *Journal of Political Economy*, 131(1):131–189. (Cited on page 4.)

Engler, Alex (2023). The ai regulatory toolbox: How governments can discover algorithmic harms. *Brookings Institution*. (Cited on page 7.)

European Union (2023). Regulation of the european parliament and of the council laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Proposal for a Regulation. (Cited on pages 7 and 8.)

Gillis, Talia B. (2021). The input fallacy. *Minn. L. Rev.*, 106:1175. (Cited on page 20.)

Gillis, Talia B and Jann L Spiess (2019). Big data and discrimination. *The University of Chicago Law Review*, 86(2):459–488. (Cited on pages 4, 17, 18, and 21.)

Goldfarb, Avi and Catherine Tucker (2012a). Privacy and innovation. *Innovation policy and the economy*, 12(1):65–90. (Cited on page 4.)

Goldfarb, Avi and Catherine Tucker (2012b). Shifts in privacy concerns. *American Economic Review*, 102(3):349–353. (Cited on page 20.)

Goldfarb, Avi and Catherine E Tucker (2024). *The Economics of Privacy*. University of Chicago Press. (Cited on page 20.)

Guerreiro, Joao, Sergio Rebelo, and Pedro Teles (2023). Regulating artificial intelligence. Technical report, National Bureau of Economic Research. (Cited on page 4.)

Guha, Neel, Christie Lawrence, Lindsey A Gailmard, Kit Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, et al. (2023). Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review, Forthcoming*. (Cited on page 4.)

Hadfield, Gillian K and Jack Clark (2023). Regulatory markets: The future of ai governance. *arXiv preprint arXiv:2304.04914*. (Cited on page 4.)

Hardt, Moritz, Eric Price, and Nati Srebro (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29. (Cited on page 4.)

Jones, Charles I (2023). The ai dilemma: Growth versus existential risk. Technical report, National Bureau of Economic Research. (Cited on pages 4 and 19.)

Kaminski, Margot E (2023). Regulating the risks of ai. *BUL Rev.*, 103:1347. (Cited on pages 7 and 9.)

Kaplow, Louis (2013). Rules versus standards: An economic analysis. In *Scientific Models of Legal Reasoning*, pages 11–84. Routledge. (Cited on page 4.)

Kaplow, Louis and Steven Shavell (1994). Accuracy in the determination of liability. *The Journal of Law and Economics*, 37(1):1–15. (Cited on page 4.)

Kiviat, Barbara (2019). The moral limits of predictive practices: The case of credit-based insurance scores. *American Sociological Review*, 84(6):1134–1158. (Cited on page 20.)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan (2018a). Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203. (Cited on page 20.)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein (2018b). Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174. (Cited on pages 4, 10, and 18.)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein (2020). Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences*, 117(48):30096–30100. (Cited on page 4.)

Liang, Annie, Jay Lu, Xiaosheng Mu, and Kyohei Okumura (2021). Algorithm design: A fairness-accuracy frontier. *arXiv preprint arXiv:2112.09975*. (Cited on page 20.)

Malgieri, Gianclaudio and Frank Pasquale (2022). From transparency to justification: toward ex ante accountability for ai. *Brooklyn Law School, Legal Studies Paper*, (712). (Cited on page 9.)

Mullainathan, Sendhil and Jann Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106. (Cited on page 20.)

NIST (2023). Artificial intelligence risk management framework (ai rmf) 1.0. Accessed: 2024-08-25. (Cited on page 10.)

Polinsky, A Mitchell and Steven Shavell (2000). The economic theory of public enforcement of law. *Journal of economic literature*, 38(1):45–76. (Cited on page 10.)

Porat, Haggai (2022). Behavior-based price discrimination and data protection in the age of algorithms. *Available at SSRN 4254326*. (Cited on page 4.)

Rambachan, Ashesh, Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig (2020). An economic approach to regulating algorithms. Technical report, National Bureau of Economic Research. (Cited on page 4.)

Rhodes, Andrew and Jidong Zhou (2024). Personalized pricing and competition. *American Economic Review*, 114(7):2141–2170. (Cited on page 4.)

Sharkey, Catherine M (2024). A products liability framework for ai. *Columbia Science and Technology Law Review*, 25(2). (Cited on page 7.)

Shavell, Steven (1984). Liability for harm versus regulation of safety. *The Journal of Legal Studies*, 13(2):357–374. (Cited on page 4.)

Shavell, Steven (1993). The optimal structure of law enforcement. *The Journal of Law and Economics*, 36(1, Part 2):255–287. (Cited on page 4.)

Shavell, Steven (2018). A model of the optimal use of liability and safety regulation. In *Economics and liability for environmental problems*, pages 77–86. Routledge. (Cited on page 4.)

Tutt, Andrew (2017). An fda for algorithms. *Admin. L. Rev.*, 69:83. (Cited on page 9.)

U.S. Senate (2023). Algorithmic accountability act of 2023. https://www.congress.gov/bill/118th-congress/senate-bill/2892. S. 2892, 118th Congress. (Cited on page 9.)