Regulating Algorithms: What and When

Talia Gillis Scott Nelson Jann Spiess

December 2024

Abstract

The regulation of algorithmic decisions, ranging from advanced credit scoring to employment screening, presents unique challenges in aligning firm behavior with regulatory goals, as well as renewed opportunities to develop the timing and methods of regulation and scrutiny. 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 analyze the tradeoffs between different regulatory regimes and 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. Using our temporal taxonomy of rule and scrutiny timing, we situate emerging and proposed AI regulations and outline an agenda for developing tools to regulate AI effectively.

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

I. Introduction

In many high-stakes domains of algorithmic decision-making, there is a misalignment between the objectives of AI system developers and deployers, and those of regulators. 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 those decisions and their impact on safety and market stability. These differences in preferences, compounded by information asymmetries and the costs of regulation, present a fundamental challenge—how should algorithmic decisions be regulated?

This chapter examines various regulatory approaches available for overseeing algorithmic decision-making, with a particular focus on the temporal design of rules and compliance scrutiny. 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 principal–agent framework, a firm (agent) develops and deploys a policy for automated decision-making—such as a lender creating a loan-underwriting model to maximize profits—while a regulator (principal) 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 game between regulator and lender plays out around two key phases. In the training phase, some initial data become available, and the firm implements its algorithm using these data. 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 be fully knowable at the training phase.

We argue that in algorithmic settings, there is a unique 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 constrain how training data can be used for building models, even before the data becomes available. For instance, restrictions on permissible inputs limit the range of models without relying on specific training data. *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 of the deployed model. These outcomes reflect both the algorithmic policy and 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 are a consequence of both the deployed algorithmic policy and the realized distribution of applicants.

The second temporal dimension concerns the timing of compliance scrutiny. 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 modeling 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.

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 realization of outcomes, which are also a function of the random deployment state that the firm has no 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 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.

To illustrate the potential value of ex-interim scrutiny, consider a financial regulator overseeing credit decisions by a lender. The regulator may, for example, be concerned with ensuring that not too many loans are given to risky borrowers. But how many such loans are given, and how many borrowers actually default, depends not only on the lender's algorithmic policy, but also on which borrowers actually apply and how the economic environment affects their repayment behavior. While the algorithm is under the lender's control, some of these additional factors are not. Scrutinizing the lender only based on realized outcomes may, therefore, be inefficient since it risks punishing the lender for bad luck rather than bad decisions. Asking the question of whether the algorithmic policy itself was reasonable may lead to a more targeted regulatory policy. Furthermore, such scrutiny is already possible before the lending rule is deployed and, thus, before any harm is done.

The desirability of ex-interim rules, however, depends crucially on the cost of scrutinizing compliance by the firm with restrictions at each stage. If scrutiny of the ex-interim stage is more resource-intensive than checking ex-post rules—for instance, because ex-post rules leverage easily measurable outcomes, while applying 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 an 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.

We connect our framework to emerging approaches in AI regulation. We begin by developing a temporal taxonomy of regulatory approaches, characterized by both the stage of the algorithmic decision-making pipeline that the rule targets and the timing at which compliance with the rule is assessed and scrutinized. This taxonomy mirrors the timing in our theoretical framework, and it enables us to explore how the full range of regulatory options has been addressed in existing regulations, proposed frameworks, and policy discussions. Within each temporal approach, we argue that regulators can adopt a variety of legal rule types, ranging from "mandates," which impose specific restrictions, to "evaluation obligations," which require firms to demonstrate that they have considered risks, evaluated alternatives, and implemented measures, such as risk-management systems, to address regulatory concerns. We consider all these legal rule types as mechanisms to constrain the algorithmic policies adopted by firms.

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). Work by Kleinberg et al. (2018) and Kleinberg et al. (2020) emphasizes the opportunities for ex-interim legal rules in algorithmic settings, where training models are determined pre-deployment. While they emphasize the potential to inspect ex-interim decisions following post-deployment disparities, 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.

Other related work has drawn on existing and emerging AI regulatory frameworks, across countries (Comunale and Manera, 2024) and 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), 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 of 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 introduces our framework, modeling regulation as a game between a firm deploying an algorithm and a regulator overseeing its use. Section III illustrate key considerations through a simple example. Section IV 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. Section V presents our taxonomy of regulatory approaches and connects the theoretical framework to current debates on AI regulation. Finally, Section VI proposes some directions for future research.

II. A Framework for Algorithm Deployment and Regulation

In this section, we model the regulation of algorithms as a game between a firm that deploys an algorithm and a regulator that oversees its use. The regulator can impose restrictions on the algorithmic policies selected by the firm, with the type of restriction depending on the stage of the decision-making timeline targeted by the regulatory rule. Rules that target the *ex-ante* stage are limited to simple restrictions, such as input restrictions, that do not rely on training data, which becomes available only at a later stage. Rules that target the *ex-interim* stage focus on restricting 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.

The regulator's choice of the timing of the rule depends on the relative costs of each approach, and whether the approach can sufficiently address the regulator's objectives. Additionally, the regulator must decide when to scrutinize compliance with a rule, which can also impact the cost of regulatory oversight. A key example is whether to evaluate compliance with *ex-interim* restrictions immediately, prior to deployment, or to defer scrutiny until after deployment.

Using our general framework, we consider several regulatory regimes and discuss additional frictions arising in the context of very complex algorithms. We then provide a simple example to demonstrate the trade-offs between different regulatory options, highlighting the opportunity for algorithmic regulation to target algorithmic decisions even before they have been deployed and possibly scrutinize compliance before harm can materialize at deployment.

II.1 Setup

We model a 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 $\mathcal{F}$. Although our framework is more general, 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 which variables are permissible to use for credit decisions or punish 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 obtain 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 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 covariates x (such as the financial histories of actual loan applicants) and some outcome of interest y (such as repayment).

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 decision function $\hat{f}_\tau \in \mathcal{F}$ based on a training signal $\tau \in T$. This signal represents the training data that is used by the firm when deciding between functions $f \in \mathcal{F}$. As a result, this stage can be thought of as running the firm's algorithm on the training data τ to come up with the algorithmic policy $\hat{f}_\tau$. 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 sample from the same distribution as the deployment. We assume that both the regulator and the firm know the joint distribution P and learn τ and θ.

Without any regulation, a firm that maximizes expected profit would choose an algorithmic policy $\hat{f}_\tau \in \arg\max_{f \in \mathcal{F}} \mathrm{E}[\Pi_\theta(f)|\tau]$ in the training phase, 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 perspective of the regulator, such as when a lender's credit-scoring policy excludes part of the target population or leads to excess

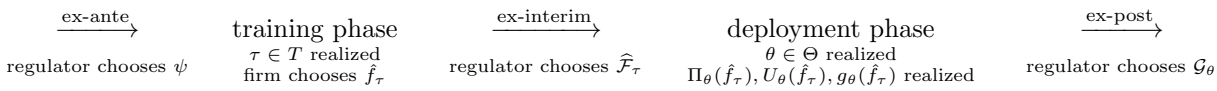

Figure 1: Rule timing of regulatory interventions.

systemic risk. The regulator, therefore, may want to restrict the firm's choice of algorithmic policy.

II.2 Target and Timing of Regulation

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 the use of 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. 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 related to the stage at which the deployment state θ is realized and the algorithmic rule is deployed, the regulator can observe some 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 punish the lender.

In all three cases, we assume that regulation defines a set of permissible algorithmic policies $\hat{f}_\tau$ that the lender can select from 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. In our model, we focus on one specific way of capturing such restrictions: we assume that these rules take the form that only specific aspects $\psi(\tau)$ of the training signal are permitted for use. This captures, for example, the scenario where training must exclude protected characteristics, $\psi(\tau)$ representing the training data after such sensitive information has been removed. That is, we represent ex-ante rules specifically as constraints applied during the pre-processing stage. Here, the choice of a rule corresponds to a choice of mapping $\psi : T \to T_\psi$, where T is the original training data space, and T_ψ represents the coarsened training signal space associated with ψ. Compliance with these rules requires that the chosen algorithmic policy $\hat{f}_\tau$ depends solely on the processed data, so that $\hat{f}_\tau = \hat{f}_{\psi(\tau)}$.[1]

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 would have been available based on the information available during

[1] As an alternative, we could think more generally of putting some coarse, easy-to-implement restrictions on the mapping $\tau \mapsto \hat{f}_\tau$. This could also include restrictions on functional forms that the lender is allowed to consider beyond pre-processing of the training data, such as ex-ante restrictions to the functional form.

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 for the realized state θ. Here, the function $g.(\cdot)$ is fixed and predefined, while the choice of the set $\mathcal{G}_\theta$ expresses the regulator's actual restriction. When there are disparity concerns, then $g_\theta(f)$ may capture realized differences in lending rates across protected groups.

Figure 1 situates each of these three rules, represented by the choices $\psi, \widehat{\mathcal{F}}_\tau, \mathcal{G}_\theta$, within the overall timeline.

We assume that determining the legal rule and compliance with it can be costly to the regulator, and that imposing ψ costs $c_{\text{ex-ante}}(\psi)$, that testing $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$ comes at a cost $c_{\text{ex-interim}}((\widehat{\mathcal{F}}_\tau)_{\tau \in T})$, and that checking $g(\hat{f}_\tau, \theta) \in \mathcal{G}_\theta$ comes at a cost $c_{\text{ex-post}}((\mathcal{G}_\theta)_{\theta \in \Theta})$. This formulation provides significant flexibility, including the ability to assign different costs to implementing regulations based on the complexity of specific rules. Specifically, the costs for testing ex-interim and ex-post rules can vary depending on the complexity of how the permissible sets are defined in relation to training and deployment data. For example, an ex-interim rule that requires that $\hat{f}_\tau \in \mathcal{F}_0$ for some *fixed* set of functions, such as simple functions, may not be costly to determine by the regulator. However, a different type of ex-interim rules that requires $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$, where $\widehat{\mathcal{F}}_\tau$ depends on the training data, may be more costly for the regulator to determine, such as checking whether disparities of the algorithmic policy are acceptable in light of disparities in the underlying training data.[2]

Rules targeting the ex-ante, ex-interim, or ex-post stages of algorithmic decisions can be scrutinized for compliance at different points in the regulatory process. For example, an ex-interim rule of the form $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$ might be scrutinized between training and deployment to ensure that only algorithmic decisions within $\widehat{\mathcal{F}}_\tau$ are implemented. Alternatively, scrutiny could occur ex-post, only if some adverse outcome materializes, to avoid expensive audits. We, therefore, distinguish between the *rule timing*—that is, which information the specific rule is based on—and the *scrutiny timing*—that is, when the rule is scrutinized. This taxonomy is discussed in greater detail in Section V below.

We assume that ex-interim scrutiny implies that the regulator forces the firm to choose a permissible function $\hat{f}_\tau$ that agree 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 of testing ex-post $c_{\text{ex-post}}$ as well as the cost of ex-interim testing $c_{\text{ex-interim}}$, 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 ψ, ex-interim rules $\tau \mapsto \widehat{\mathcal{F}}_\tau$, ex-post rules $\theta \mapsto \mathcal{G}_\theta$, and the decision when to scrutinize each).[3]

[2] 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 since it is performed only when bad outcomes materialize, thus reducing the expected cost. A natural extension would be to explicitly add different costs based on scrutiny timing as well.

[3] 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.

II.3 Regulatory Regimes

We now make these different regulatory options concrete by listing five specific regulatory regimes. We consider the case where ex-ante rules can be scrutinized for free, $\bar{c}_{\text{ex-ante}} = 0$, which expresses the idea that restrictions on pre-processing may be easy to verify. Ex-interim rules cost $\bar{c}_{\text{ex-ante}}$ to enforce, no matter the complexity of the rule.[4] Finally, we assume that inspecting outcomes comes at a flat cost of $\bar{c}_{\text{ex-post}}$. Typically, we would assume that ex-post rules are cheap to enforce while checking ex-interim rules may be more costly ($\bar{c}_{\text{ex-interim}} > \bar{c}_{\text{ex-post}} \geq 0 = \bar{c}_{\text{ex-ante}}$) 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. 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. If the regulator imposes *ex-ante rules* ψ, the firm is forced to choose $f \in \mathcal{F}$ based on $\psi(\tau)$ only, so $\hat{f}_\tau = \hat{f}_{\psi\tau}$.

3. If the regulator performs an *ex-interim audit*, the regulator pays $\bar{c}_{\text{ex-interim}}$ and chooses $\widehat{\mathcal{F}}_\tau$, and the firm chooses $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$.

4. 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 $\bar{c}_{\text{ex-post}}$ 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 $U_\theta(\hat{f}_\tau) - \bar{c}_{\text{ex-post}}$ since the algorithmic policy is already deployed.

5. 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 $\bar{c}_{\text{ex-post}}$, whether to perform an in-depth audit. If the rule is violated, the regulator pays an additional $\bar{c}_{\text{ex-interim}}$ and chooses $\widehat{\mathcal{F}}_\tau$. 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 $\bar{c}_{\text{ex-post}} + \delta\bar{c}_{\text{ex-interim}}$.

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

II.4 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 use deep neural networks or other machine learning methods to decide who should get 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)

[4] An alternative would be to distinguish between simple and complex restrictions. Simple restrictions would be of the form $\hat{f}_\tau \in \mathcal{F}_0$ where the permissible set $\mathcal{F}_0$ does not depend on the training data, at cost $\bar{c}^{\text{simple}}_{\text{ex-interim}} \geq 0$. This would include restrictions on the functional form that is separate from the actual coefficients. Complex restrictions would be of the more general form $\hat{f}_\tau \in \widehat{\mathcal{F}}_\tau$ at cost $\bar{c}^{\text{complex}}_{\text{ex-interim}} > \bar{c}^{\text{simple}}_{\text{ex-interim}}$, where the allowable functions themselves depend on the data.

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 capture the restriction that the regulator may not be able to fully capture complex AI algorithms by restrictions of the form

$$\mathcal{F}^* = \{f \in \mathcal{F}; \phi(f) \in \Phi^*\}.$$

That is, rather than being able to distinguish between all functions $f \in \mathcal{F}$, the regulator can only distinguish between functions based on simpler descriptions $\phi(f)$. Concrete examples from our previous work include audits based on simple explainers (Blattner, Nelson, and Spiess, 2024) and "discrimination stress testing" based on evaluating an algorithmic policy by evaluating it at specific test points (Gillis, 2021).

III. 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 as our firm, 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 highly 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 in the low state. Neither the regulator nor the lender prefer the imprecise rule. Concrete profits and regulator utility that represent these rank-orderings are provided in Table 1.

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

Table 1: Lender and regulator payoffs in the example.

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

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 $\tau = \text{likely high}$, then regulator and lender agree that the risky algorithmic policy is optimal. But if $\tau = \text{certainly low}$, then the regulator prefers the conservative policy, while the lender still prefers the risky policy.

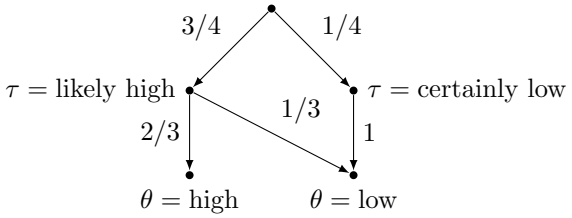

Figure 2: Joint distribution of training and deployment signals in the example.

We now consider the different regulatory regimes from Section II.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 $\bar{c}_{\text{ex-interim}} = c$ with $0 < c < 3$, and that ex-ante and ex-post tests are not costly ($\bar{c}_{\text{ex-ante}} = 0 = \bar{c}_{\text{ex-interim}}$). The different regulatory regimes then play out as follows:

1. 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. If the regulator can only impose *ex-ante rules* of the form $\hat{f}_\tau = \hat{f}_{\psi\tau}$, then forbidding the lender to distinguish between the two training states means that the lender still makes the risky choice.

3. 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 ($g_\theta(\hat{f}_\tau) = \{bad\}$). 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. If the regulator *intervenes at the interim stage*, they can pay a fee of $\bar{c}$ 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 $2.75 - \bar{c}$, 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. 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 $\bar{c}$ to check whether the lender chose $\hat{f}_\tau$ = risky for τ = certainly low, but only in the case that θ = low and $\hat{f}_\tau$ = 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 $\bar{c}/4$, 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 $2.75 - \bar{c}/4$. Expected lender profit is unchanged at 3.75.

Of course, the results would change with different parameters and variants of the model. For example, assume that the regulator could impose simple restrictions $\hat{f}_\tau \in \mathcal{F}_0$ of functional forms at a lower cost (or no cost at all), as in Footnote 4. In this case, the regulator could rule out the risky function altogether. With our parameters, this would lead to safe regulator utility of 2 (but reduce expected lender profit to 2.5). The regulator could, however, prefer this option over the other alternatives if the costs of alternative restrictions are high and the regulator is very prudent and wants to rule out the bad outcome altogether.

In the stylized example, so far, we have not considered any complexity constraints of the type from Section II.4. Intuitively, such frictions would 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). To provide a concrete example, we 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.

IV. General Implications for Regulating Algorithms

The prior section introduced a model 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 of the model and discuss them in connection with other related work.

IV.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 model, 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 ($\bar{c}_{\text{ex-interim}} = 0$), 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 $\hat{f}_\tau = \hat{f}_{\psi\tau}$. 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 II 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.

IV.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 $\bar{c}_{\text{ex-interim}}$ of accessing the training data—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 V below points to some real-world examples of such interventions, and Section III provides a theoretical illustration.

To build intuition for this result, an analogy from a classic example in the economics of regulation may be helpful. Consider the case of cost-of-service regulation (e.g., Cicala, 2022). In cost-of-service regulation, an agent (e.g., a price-regulated electric utility) is compensated with revenue that is approximately affine in specific components of their per-unit cost structure, such as capital and fuel expenses, assuming demand is close to price-inelastic. This incentive scheme penalizes the agent for other cost realizations that are less observable and verifiable, such as expenses related to certain forms of research and development, which are not factored into the compensation rule. Unlike traditional agents, who cannot be effectively regulated at the ex-interim stage by promising a certain level of profits (as opposed to revenue) in return for incentive-aligned choices at the interim stage due to the difficulty of observing and verifying their actions, an *algorithmic* agent offers an opportunity for ex-interim regulation because its processes and decisions—such as model development and data usage—can, in theory, be observed by a regulator. However, if scrutinizing ex-interim decisions is costly for the regulator, it may be optimal to only scrutinize the ex-interim decision in the case of particularly adverse ex-post outcomes; in the case of a regulated electric utility, this could be a high price of electricity in the bad state of the world where expensive research and development did not yield the cost savings that the utility anticipated.

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.

IV.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 model, ex-ante restrictions refer to limitations on algorithmic restrictions that do not rely on any training or deployment data. This ex-ante regulation can take several forms. In our model, they take the form of pre-processing of the data, and capture the ex-ante exclusion of specific inputs before training happens. In a broader sense, we could think of simple ex-ante restrictions as also including restrictions to simple functional forms.

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)[, presented at this conference,] 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).[5]

IV.4 Other Considerations

Our framework necessarily excludes some features that may also be relevant for the context-specific regulation of algorithms. One important consideration not included above is the demand for privacy and other normative

[5] 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.

considerations in algorithmic regulation. Ex-ante restrictions on certain 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 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.[6] 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. Especially questions of fairness can be as much about which data is used as about how the algorithm is trained, and obtaining high-quality data that represent the deployment distribution well may be a costly investment that firm and regulator have different preferences over.

V. A Taxonomy of Emerging Regulatory Approaches

In this section, we examine how our framework informs and connects to emerging regulatory strategies for AI and ongoing policy debates. Our earlier discussion highlighted how regulators face choices over the stage of intervention within the lifecycle of algorithmic development or deployment—and the distinctive opportunities presented by algorithmic settings. Here, we build upon these stages to form a taxonomy of regulatory approaches and show how examples from current discussions around AI governance fall within this scheme.

Our taxonomy is based on the distinction between "rule timing" and "scrutiny timing", which we now make more explicit. By "rule timing," we refer to the phase in the development or deployment pipeline that a rule is designed to govern. In contrast, "scrutiny timing" pertains to the stage at which compliance with that rule is assessed—whether before training, after training and before deployment, or after deployment. This distinction underscores the temporal choices regulators must navigate: not only when to impose rules but also when to

[6] For an interesting related analysis, see work in Callander and Li (2024)[also presented at this conference].

evaluate adherence to them.

Table 2 illustrates the temporal choices available to regulators. While ex-ante rules—such as input restrictions—can theoretically be scrutinized at any of the three stages (e.g., as part of a pre-development licensing process, monitored ex-interim, or evaluated only after deployment), other rule timings impose more limited scrutiny options. For example, an ex-interim rule, such as a requirement on the model's fit to the training data, can only be scrutinized during or after the ex-interim stage (e.g., at the ex-interim or ex-post stage), but not ex-ante, since scrutiny of the relationship of the model to training data cannot occur before the training data is available.

Scrutiny timing → Rule timing ↓	**ex-ante** (before training)	**ex-interim** (between training and deployment/during development)	**ex-post** (after deployment)
ex-ante (before training)	Legal rule targets the ex-ante stage and is scrutinized ex-ante. Example: input restrictions (ex-ante rule) that are required for a license (ex-ante scrutiny).	Legal rule targets the pre-development stage but is only scrutinized at the development stage. Example: demonstrating compliance with an input restriction (ex-ante) through a submission of a conformity assessment before model deployment (ex-interim)	Legal rule targets pre-development stage but is only scrutinized ex-post. Example: demonstrating compliance with an input restriction (ex-ante) through a post-hoc analysis following a claim that a model caused unfair outcomes for a minority group.
ex-interim (between training and deployment/during development)	N/A	Legal rule targets development stage and is scrutinized pre-deployment. Example: an explainer of model weights (ex-interim) that must be documented and reported to the regulator.	Legal rule targets development stage but is only scrutinized at deployment. Example: Evaluation of whether inappropriately trained model parameters (ex-interim) were the cause of disparities at deployment (ex-post).
ex-post (after deployment)	N/A	N/A	Legal rule targets outcomes and is scrutinized at deployment. Example: Deployment impact assessments (ex-post) showing the true impact on different groups.

Table 2: Temporal taxonomy of regulatory options.

It is important to recognize that within each rule–scrutiny temporal choice, regulators can deploy a wide variety of rule types. Policy discussions around the regulation of AI generally fall into two broad categories. The first category comprises *mandates*, where legal rules prescribe specific conduct. Mandates can take the form of clearly defined directives, such as input restrictions or requirements to conform to specified standards, which are often referred to in the law and economics literature as "rules" (Kaplow, 2013). Alternatively, mandates may include more flexible requirements that, while subject to interpretation, are primarily designed to enforce compliance with specific obligations (e.g., ensuring the representativeness of training datasets). These types of

standards, though adaptable, function closer to mandates because their focus is on achieving compliance with a predetermined requirement.

The second category, which we term *evaluation obligations*, covers legal requirements aimed at demonstrating deliberation and risk consideration rather than prescribing specific actions. Evaluation obligations typically require actors to engage in processes like conducting assessments, evaluating risks, or documenting decision-making without dictating particular outcomes. In the law and economics framework, these are often closer to "standards," as their interpretation depends on the specifics of the context and is often assessed ex-post. For example, a rule requiring a developer to identify and mitigate foreseeable risks through a documented process would fall under evaluation obligations, as its focus is on promoting transparency, accountability, and iterative risk management.[7]

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 2. Relative to the formal model in Section II, we apply the three temporal stages and possible regulatory interventions more broadly in this section to encompass a wider range of instruments. While our focus is on scrutiny through public enforcement, private enforcement, such as private litigation, can also play a role in ensuring compliance; however, this scrutiny typically occurs ex-post, following the materialization of harm. Below, we examine the various choices of rule timing and how these rules can be scrutinized at different stages of algorithmic development and decision-making.

V.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. Within the model, we considered a particular type of ex-ante restriction, where only a specific aspect of the training signal $\psi(\tau)$ can be used by the firm. This would include an ex-ante legal *mandate* of input restrictions, which prohibit certain types of information from being provided to a model during its development. The EU AI Act (European Union, 2023), for example, contains several provisions related to input restrictions, such as Article 5(1)(c), which restricts the use of information on "social behavior or known, inferred, or predicted personal or personality characteristics" for uses not in the context in which the information was collected. Another example in the EU AI Act is the prohibition on the use of emotion recognition in workplaces or educational institutions (except for medical and safety reasons). Within our model in Section II, this are restrictios on the aspects of the training signal available to the firm (mandate), which can be imposed on firms already at the ex-ante stage before the particular training state τ is known.

Another type of ex-ante rule, not captured explicitly by our model's definition of ex-ante rules and more in the spirit of Footnote 1, could be an *evaluation obligation* that requires model developers to assess and document potential risks associated with the type of algorithmic policy they intend to create. These obligations aim to ensure that risks are considered at the earliest stages of the pipeline, to allow for risk mitigation before development

[7] Although this distinction can be helpful in considering the different ways in which requirements operate, the distinction between them may be blurred at times. For example, a legal rule may require the establishment of a risk management system. If the obligation specifies how the system should be set up, such as defining required steps, documentation protocols, or adherence to certain standards, it aligns more closely with a *mandate* (even if it is a flexible standard). If the obligation focuses on the process of identifying and managing risks, leaving the specifics of implementation open to interpretation, it is better classified as an *evaluation obligation*. In this case, the rule emphasizes deliberation and ongoing assessment rather than compliance with a predefined structure. Article 9 of the EU AI Act, which requires the establishment of a risk management system considering foreseeable risks to "health, safety or fundamental rights," would likely fall into the *evaluation obligation* category, as it focuses on process-oriented compliance.

begins. Below we consider how both ex-ante *mandates* and *evaluation obligations* 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, as discussed in Section II.

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 (mandate) or adequately addressing potential risks through documented risk assessments (evaluation obligation).

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 (mandate) or requirements to identify and mitigate risks at the pre-development stage (evaluation obligation). 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) suggest 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 particularly high, as discussed in Section II, 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 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. 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.

V.2 Ex-Interim rules

Ex-interim rules apply to the model development stage, after training data is available but before deployment, and therefore relate to rules that restrict or constrain the algorithmic function once the training state τ is known. At this point in time, the regulator can impose rules that relate to the training signal and the concrete function trained on that data. As described in Section II.2, ex-interim rules can be restrictions on the function $\hat{f}_\tau$ and may depend on the training data itself. For instance, a regulator might assess whether a better algorithmic rule could have been implemented based on the information available during training, using tools such as explainers or stress-testing frameworks, as discussed in Section IV.1.[8] For example, Article 9 of the EU AI Act requires a risk management system that considers the foreseeable risks to "health, safety or fundamental rights" at deployment (evaluation obligation). 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.

Another type of ex-interim requirement puts restrictions on the training dataset—which can be understood as targeting the training signal τ in an extension of our framework to the case where the firm also has some control over τ—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.[9] Sometimes ex-interim rules take the form of evaluation obligations, which indirectly constrain the particular model firms are able to adopt.

However, 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 take place. Therefore, scrutiny of ex-interim rules typically occurs during the ex-interim or ex-post stages.

Ex-interim scrutiny. Ex-interim scrutiny refers to 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 ex-interim.

[8] 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.

[9] While this requirement might seem somewhat vague and open to interpretation, we classify it as an ex-interim *mandate* rather than an *evaluation obligation*, as it imposes restrictions on developers during the model development stage rather than a requirement to generally assess risks.

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 the ex-interim rule after the deployment of the AI system. Where regulators are risk neutral relative to firms, it may be beneficial for regulators to only scrutiny ex-interim compliance after there is ex-post materialized harm or some other indication of lack of compliance to selectively target enforcement efforts to the most concerning incidents.

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,[10] the primary consideration of 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.

V.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, assessing harm that has, at least partially, already occurred. As discussed in Section II, the materialized harm is a function of the model $\hat{f}$ and training signal τ, but also of the eventual state of the world θ, which is only known ex-post. For instance, a regulator could observe outcomes such as lending disparities, and impose sanctions on the lender.

One common form of a *mandate* ex-post rule is tied to outcome-based obligations, such as ensuring that an AI system does not produce discriminatory outcomes or safety-critical failures in deployment. For example, tort liability frameworks may hold developers or deployers accountable if a deployed AI system causes physical harm or other damage, such as an autonomous vehicle involved in a traffic accident. Particularly in cases of strict liability, where showing that the deployer took reasonable steps to reduce risk does not absolve liability, the firm becomes a de facto insurer against harm.

Evaluation obligations relating to monitoring, measuring, and reporting outcomes frequently target the ex-post stage. For example, the EU AI Act in Article 79 requires post-market monitoring, in light of real-world outcomes, creating an *asessment obligation* ex-post. Similarly, the proposed Canadian Act (Canada, 2024) includes provisions for monitoring and reporting incidents of harm or biased output resulting from the use of high-impact AI systems. The NIST AI Risk Management Framework (NIST, 2023), a non-binding framework in the U.S. for managing risks associated with AI systems, requires that system performance at deployment be compared to the

[10] 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.

assessments made pre-deployment.

As noted earlier, 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. However, when firms are more risk-averse than policymakers, ex-post rules may lead to overly conservative behavior that is not socially optimal, particularly since outcomes depend on the deployment state θ, which is beyond the firm's control. 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).

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 already 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 exciting 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 how should incentives for those be designed?

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 page 15.)

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb (2023). Do we want less automation? *Science*, 381(6654):155–158. (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 and 18.)

Blattner, Laura, Scott Nelson, and Jann Spiess (2024). Unpacking the black box: Regulating algorithmic decisions. *arXiv preprint arXiv:2110.03443*. (Cited on pages 4, 9, 11, 12, 14, and 21.)

Callander, Steven and Hongyi Li (2024). Regulating an innovative industry. Technical report, National Bureau of Economic Research. (Cited on pages 4 and 15.)

Canada (2024). Artificial intelligence and data act (aida). Proposed Legislation. (Cited on pages 19 and 20.)

Cicala, Steve (2022). Imperfect markets versus imperfect regulation in us electricity generation. *American Economic Review*, 112(2):409–441. (Cited on page 13.)

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.)

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 17 and 18.)

Gillis, Talia B. (2021). The input fallacy. *Minn. L. Rev.*, 106:1175. (Cited on pages 9 and 14.)

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, 12, 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 15.)

Goldfarb, Avi and Catherine E Tucker (2024). *The Economics of Privacy*. University of Chicago Press. (Cited on page 15.)

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 14.)

Kaplow, Louis (2013). Rules versus standards: An economic analysis. In *Scientific Models of Legal Reasoning*, pages 11–84. Routledge. (Cited on pages 4 and 16.)

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 15.)

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 14.)

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, 12, and 20.)

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 14.)

Mullainathan, Sendhil and Jann Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106. (Cited on page 14.)

NIST (2023). Artificial intelligence risk management framework (ai rmf) 1.0. Accessed: 2024-08-25. (Cited on page 20.)

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 21.)

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.)

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.)