Incentive Aware AI Regulation

Anonymous Author(s)
Affiliation
Address
email

Abstract

The EU AI Act emphasizes the importance of differentiated safety requirements across classes of users. However, machine learning (ML) service providers may strategically under-enforce such requirements to reduce development costs or accelerate deployment. We study this phenomenon through the lens of a principal–agent model, where regulators act as principals enforcing risk-control obligations, while ML service providers act as agents with private incentives. A key challenge is that direct enforcement of safety constraints is often infeasible, since verification requires costly monitoring and statistical uncertainty may be exploited by strategic agents. To address this, we introduce incentive aware statistical protocols—rules tailored for the providers given their private costs, that translate observed model performance into enforceable outcomes, such as licensed market access. We show that these protocols can be designed to guarantee obedience to regulations: providers who do not comply with user-specific safety requirements are statistically driven to self-exclude from the market, while compliant providers remain viable. Our framework provides new theoretical insights into the intersection of statistical testing, mechanism design, and trustworthy AI regulation, offering a foundation for the development of enforceable AI governance mechanisms.

1 Introduction

Machine learning models have achieved remarkable empirical success across many domains such as language modelling [Vaswani et al., 2017] and image generation [Rombach et al., 2022]. With these promising results, we now see widespread real-world applications of machine learning models such as credit scoring [Baesens et al., 2003], social justice [Angwin et al., 2022] and other high-risk applications. Despite these successes, the risk assessment of blindly relying on the predictions of these models is considered catastrophic [Voigt and Von dem Bussche, 2017, Veale and Zuiderveen Borgesius, 2021, Laux et al., 2024] as much of these models typically fail to achieve OOD generalisation [Sagawa et al., 2020, Eastwood et al., Singh et al., 2024] or are vulnerable to adversarial attacks [Szegedy et al., 2013, Goodfellow et al., 2014] and often lack the notions of fairness across different subgroups [Chouldechova and Roth, 2018, Mehrabi et al., 2021, Barocas et al., 2023]. In light of these limitations of modern AI methods there have been recent attempts by the policy makers to regulate the real world applications of these AI models, including the EU AI act [Edwards, 2021], which is one of the most comprehensive regulations spanning across the range of AI applications, recommending differentiated safety requirements and user-specific risk control.

However, in practice regulators are confronted with unverifiable black boxes, costly monitoring, and pervasive statistical uncertainty that firms can exploit to under-enforce safety [De Almeida et al., 2021], creating a fundamental enforcement tension: economic incentives of providers interact with noisy verification to produce strategic non-compliance. Technical responses address parts of this space: with differential privacy [Dwork et al., 2014], certified-robustness methods for train [Cohen et al., 2019] and test time [Seferis et al., 2023, Singh et al., 2023]; complementary work on documentation

Submitted to 1st PAIG Workshop at EurIPS 2025, Copenhagen. Do not distribute.

and auditability (e.g., model cards) aims to improve transparency for private and self oversight [Mitchell et al., 2019], while empirical audits expose real-world distributional harms [Buolamwini and Gebru, 2018]. However, in practice, the model designers often have strategic incentives to game regulations by manipulating the statistical uncertainty in their favour. Verification, auditing or benchmarking solutions [Jansen et al., 2024, Li and Goel, 2025, Hardt, 2025, Arias et al., 2025] to regulations often ignore this strategic aspect. Economic and policy scholarship has identified limitations of single entity regulatory frameworks and thus propose idea of *private regulation* [Ball, 2025]. Private regulation encourages private or self regulating mechanisms for AI governance [Stein et al., 2024] allowing for regulatory market design [Hadfield and Clark, 2023, Tomei et al., 2025].

These cross-disciplinary efforts have now laid the foundations and normative ideals of private regulation but leave open how to design enforcement rules that are resilient to sampling noise and strategic actors, motivating approaches that treat verification and rulemaking as a unified mechanism under statistical uncertainty. Inspired from the recent developments in incentive aware hypothesis testing [Bates et al., 2022, 2023, Min, 2023, Hossain et al., 2025] we propose a principal-agent framework focusing on the outcome/incentives rather than decisions to regulate AI model providers via statistical testing. Our framework consists of a principal (public or private regulator) who rather than trying to solve a decision problem of who satisfies regulations or not via statistical testing ensures that the statistical test is aware of the welfare objectives of the regulator and the incentives of the model designers. The awareness of model designer's incentives while ensuring compliance to regulations allows the statistical test to keep null agents out of the market without being too strict, thus encouraging maximum market participation even when prior population of null agents is much larger than agents who abide by the regulations. Additionally we characterise the regulation enforcing incentive aware tests via desirable gambles and provide the sufficient conditions under which incentive aware tests could encourage model designers to improve their models. We also discuss the scenarios where model designers costs are their private information and they might lack precise knowledge about the properties of their model for them to effectively strategise using it.

2 Preliminaries

Why regulating an AI model is relatively hard. Regulations are ubiquitous to almost everything that humans use in the real world, from physical goods to processes. Since AI has started to become more and more useful in the real world, concerns on regulating its ill-effects have also become important. However, regulating AI is slightly different from classical regulations. Although classical regulations had to deal with statistical uncertainty, AI regulation has brought it to the center of the discussion. In a classical scenario, where we want to regulate a physical good, checks for physical goods are practically deterministic. An example of a check on a physical good could be as follows. We weigh an apple and compare the weight to a threshold. Process-level regulations for apple production introduce aleatoric uncertainty because apple instances vary; a regulator could propose a test about the apple production process as a statement that the average weight of apples μ is less than a threshold μ_0, i.e. $H_0 : \mu \geq \mu_0$, by sampling items and using classical inference on the population mean.

Machine learning adds a second stochastic layer for regulation via statistical inference. A training procedure is a learning algorithm $\mathcal{A} : \mathfrak{D} \times \Xi \to \mathcal{H}$, which takes in a dataset $D \in \mathfrak{D}$, a hyperparameter $\xi \in \Xi$ and selects a model from the hypothesis class $\mathcal{H}$. Since the data-generation mechanism construes the dataset D as random, the algorithm defines a distribution over models $\mathcal{A}(D, \xi)$, where ξ is a hyperparameter which can vary. This captures the epistemic uncertainty due to the finite sample training, where changing the dataset D during training, changes the final model. Thus each model $h \in \mathcal{H}$ has a risk, which is itself a random variable $r(h)$ evaluated on the entire population. In practice it can only be estimated with finite number of samples. Thus a natural instance-level requirement $H_0^{\text{inst}} : \ r(h) \leq \tau$ is stochastic in nature. While a process-level hypothesis H_0^{proc}, for example

$$H_0^{\text{proc}} : \ \Pr_{D,\xi} \left(r(\mathcal{A}(D, \xi)) \leq \tau \right) \geq p$$

has two levels of stochastic uncertainty. Certifying H_0^{proc} requires sampling across datasets, seeds, and evaluation draws; certifying H_0^{inst} requires tight estimation bounds on $r(h)$. The two nested sources of randomness (model generation and model evaluation), together with non-i.i.d. data, distribution shift, and adversarial inputs, make statistical error both larger and structurally different from manufacturing variability. That structure creates real gaps for regulation: finite samples and misspecified benchmarks yield ambiguous outcomes and provide actors with plausible statistical deniability that a process "did nothing wrong". Not even going into the the regulation of an algorithm, this statistical uncertainty also challenges the regulation design at model level. For example, legal and definitional gaps make it hard to say precisely does a "model" incur a legally actionable "harm" [Kroll, 2015, Edwards, 2021]. We argue that the inability to define a precise requirement is also because of this statistical uncertainty. After all, if a requirement deems acceptable for AI models to be no harm, if they cause harm with very low probability, who indeed are we okay with being subject to these unfortunate events? Since statistical statements assume individual subjects as exchangeable as they them come from the same underlying population. However, individuals are interested in their unique individual harm which relatively harder to define statistically. Beyond these statistical issues, AI model regulation confronts a dense web of non-statistical obstacles that do not arise, or arise far less severely, for physical goods (See Appendix A.5).

Imprecise Probability, Gambles and E-values Standard probability theory assigns a unique numerical value to each event, whereas *imprecise probabilities* (IP) allows a range of plausible values to represent uncertainty in the presence of limited or ambiguous information [Walley, 1991, Troffaes, 2007, Augustin et al., 2014, Troffaes and de Cooman, 2014]. This extension of classic probability theory comes from the subjective interpretation of probability [de Finetti, 1974]. Central to the subjective interpretation is a notion of a gamble. A gamble is a bounded real-valued payoff function whose fair price reveals an agent's subjective probability; this betting interpretation underlies the operational meaning of probability. Formally we represent a gamble as $G : \Omega \to \mathbb{R}$ where Ω usually refers to the sample space of a probability measure. Gambles as pay-offs of an uncertain event enjoy a close relationship to actuarial risk and insurance. Gambles are also closely connected to the recent developed in game-theoretic statistics and hypothesis testing, called e-values [Derr and Williamson, 2024]. An e-variable (whose observed value is often called an e-value) is a non-negative random variable E with $\mathbb{E}_P[E] \leq 1$ for all $P \in \mathcal{P}_{null}$. E-values quantify evidence against a null in expectation and admit interpretation as the outcome of a fair bet or wealth of a skeptic. They are naturally composable under optional continuation, and provide a practical alternative to p-values for sequential any-time valid testing [Vovk and Wang, 2021, Shafer, 2021, Ramdas et al., 2023, Grünwald et al., 2024]. Together these notions connect betting-style evidence, robust representations of uncertainty, and evidence-based tests in a way that is directly applicable to the two-layer statistical problems that arise in AI regulation: they allow regulators to formalize the incentives on the evidence than formalising the regulation as a statistical decision-making ambiguity task, allowing connection between statistical testing and mechanism design, also guiding how different kinds of sample-based evidence can be combined to enforce the desired properties.

3 Incentive Aware Regulation

We consider $\mathcal{X} \subseteq \mathbb{R}^d$ as our instance space and $\mathcal{Y}$ as our target space, where for regression tasks $\mathcal{Y} \subseteq \mathbb{R}$, and for a K-class classification problem $\mathcal{Y} = \{1, \ldots, K\}$. We consider a supervised learning scenario where $\mathcal{H}$ denotes the hypothesis class of functions $f : \mathcal{X} \to \mathcal{Y}$. In our incentive aware regulation framework we consider two agents (1) Model Designer and a (2) Regulator along with nature. We assume that nature typically reveals an $x \in \mathcal{X}$ and then later a corresponding $y \in \mathcal{Y}$ is also revealed. We assume that the nature is stochastic, i.e., there exists a fixed but unknown distribution $P(X \times Y)$ where X and Y denote the random variables on $\mathcal{X}$ and $\mathcal{Y}$ respectively. We also assume a fixed loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ which quantifies the risk of a prediction of a model.

Definition 3.1 (Requirement). Let $\mathfrak{R} : \mathcal{H} \times \Delta(X, Y) \to \{0, 1\}$ be a requirement. A model f is regulation compliant with respect to nature P and requirement $\mathfrak{R}$ in deployment if $\mathfrak{R}(f, P) = 1$.

A concrete instantiation of Definition 3.1 is a ϵ-safety regulation, i.e., model f whose expected risk with respect to nature under a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is controlled by ϵ. Mathematically, $\mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)] \leq \epsilon$, then regulation $\mathfrak{R}_{\epsilon,\ell}$ for an ϵ-safe model f is $\mathfrak{R}_{\epsilon,\ell}(f, P) = 1$ and 0 otherwise. The regulation divides the set of machine learning models into two categories, null and alternate. The models belonging to null do not satisfy the regulation i.e. $\mathcal{H}_0 := \{f : \mathfrak{R}(f) = 0\}$ and alternate i.e. $\mathcal{H}_1 := \{f : \mathfrak{R}(f) = 1\}$, also $\mathcal{H}_0 \cap \mathcal{H}_1 = \emptyset$ and $\mathcal{H}_0 \cup \mathcal{H}_1 = \mathcal{H}$. We assume that the knowledge of whether the deployed model satisfies the regulation or not is private to the model designer. In practice the model designer may not exactly know if the model will satisfy the regulation, however, they typically have much more information about the model than the regulator.

This additional knowledge serves as private information of the model designer. Thus we define the evidence $Z := \ell(f(X), Y)$ and the push forward measure $P(Z)$ as the distribution of risks dependent on the model f and we define the set of all risk distributions as $\mathcal{P}$ with, naturally, $P \in \mathcal{P}$.

3.1 Incentive Aware Statistical Protocol

An incentive aware statistical protocol is a menu of statistical contracts $\Pi := \{\pi : \mathbb{Z} \to \mathbb{R}_{\geq 0}\}$ which in practice act as licenses that the model providers can obtain from the regulator to earn profit $\pi(Z)$ given the statistical evidence of their model $Z \in \mathbb{Z}$. Ideally, we would want to ensure the following property in the statistical contracts offered to the provider. Obedience to the regulation, i.e., the agents that do not fulfil the regulation are incentivised to self-exclude from the market and incentive compatibility for the obedient agents. In the example of ε-saftey regulation, the more an agent invests effort in training representative models that incur lower risk, the larger market share license is available to them.

Definition 3.2 (Obedience to regulation)**.** A menu of license contracts Π is said to enforce obedience to regulation if the following holds true ex-ante for the agents

$$\sup_{\pi \in \Pi} \mathbb{E}_{Z\sim P}[\pi(Z)] \leq C \quad \forall P \in \mathcal{P}_0$$

where C is the overall market entry fee for all the model designers. Obedience definition 3.2 ensures that the model designers who are violating the market regulation can not recover their entry fee from any of the market licenses $\pi \in \Pi$. In other words we cap the profits for the non-obedient model designers that their cost of operating in the market makes them self exclude from the market. Using tools from the theory of desirability [Augustin et al., 2014, Walley, 1991], we can characterise the licensing menu Π. A gamble is $G : \Omega \to \mathbb{R}$ where Ω refers to the sample space of a probability measure, in our case $\Omega = \mathbb{Z}$ denoting the space of all the possible values evidence Z can take.

Definition 3.3. (Set of Desirable Gambles) A set of gambles $\mathfrak{G} = \{G | G : \Omega \to \mathbb{R}\}$ is desirable with respect to $\mathcal{P}_0$ if $\inf_{P\in\mathcal{P}_0} \mathbb{E}_P[G] > 0 \quad \forall\, G \in \mathfrak{G}$. The set $\mathfrak{G}_{\geq 0}$ is called marginally desirable if above inequality is not strict.

Proposition 3.4. *A menu of license functions* Π *induces obedience to regulations if and only if* $C - \Pi \subseteq \mathfrak{G}_{\geq 0, \mathcal{P}_0}$*, where* $\mathfrak{G}_{\geq 0,\mathcal{P}_0}$ *is the set of all marginally desirable gambles with respect to* $\mathcal{P}_0$*.*

The above result characterises the menu of license functions Π that can enforce obedience to regulations. A menu of licenses enforces obedience if and only if the gambles $C - \Pi$ induced by it are desirable to the regulator with respect to the set of distributions $\mathcal{P}_0$. A useful consequence of this characterisation is that the regulator, once they know the market entry fee they charge a provider and offer desirable gambles from their perspective as menu for the model designers. Additionally, we define a preference relationship $\succ$ on the space of evidences $\mathbb{Z}$ as

Definition 3.5 (Evidence Preference)**.** A regulation requirement $\mathfrak{R}$ induces an incomplete preference relation ($\succ_{\mathfrak{R}}$) on the space of evidence $\mathbb{Z}$.

The above definition states that in light of a regulation requirement a regulator has a preference for evidence i.e. assume that $Z_1, Z_2 \in \mathbb{Z}$ are evidences generated by models f_1 and f_2 respectively, and $\mathfrak{R}(f_1) = 1$ while $\mathfrak{R}(f_2) = 0$ respectively. Then $Z_1 \succ_{\mathfrak{R}} Z_2$. Assuming the space of evidence $\mathbb{Z}$ has a natural total order $\succ$, then the $\succ_{\mathfrak{R}}$ must agree with this total order. For example, let's assume the evidence to be some loss value i.e. $Z = \ell(f(x), y) \in \mathbb{R}_{\geq 0}$, then the total order for loss would be $Z_1 \succ Z_2$ if $Z_1 < Z_2$, i.e. lower the loss the better the evidence. This allows us to discuss a second desirable property of a menu Π. Ideally, from the perspective of the model designer who is presented with a menu of contracts, they must not be penalised for improving upon their evidence. We call such menu incentive compatible from the perspective of the model designer. A benevolent regulator would also encourage improvement of the evidence from the model designers thus aligning, their incentives. Formally we define incentive compatibility as

Definition 3.6 (Incentive Compatibility of the Menu)**.** Assuming a designer can make two models f_1 and f_2 such that they produce risks distributions P_1 and P_2 such that $\mathbb{E}_{Z\sim P_1}[Z] \succ \mathbb{E}_{Z\sim P_2}[Z]$ then a menu of licenses is incentive compatible if

$$\max_{\pi\in\Pi} \mathbb{E}_{Z\sim P_1}[\pi(Z)] > \max_{\pi\in\Pi} \mathbb{E}_{Z\sim P_0}[\pi(Z)]$$

Definition 3.6 states that for a model designer who invests effort in improving their model and thus provides better evidence on average, the licensing function must ensure a better expected revenue for that agent and hence present an incentive to strive for making a better model. An incentive compatible menu Π by controlling the incentives based on the outcomes encourages self-governance. We also discuss that the monotonicity of the menu in evidence total order is sufficient to ensure self governance.

Proposition 3.7. *A menu Π is incentive compatible according to Def 3.6 if for all $\pi \in \Pi$, π is monotone in total order on $\mathbb{Z}$.*

Model Designer's Optimal Response In this section we describe the model's designer behaviour model that we consider as part of analysis in the principal agent formulation of our problem. While prior works in principal agent hypothesis testing assume the agent to be perfectly informed about their type, such assumptions are too strong in our setting for model designers as agents as they may have more information than the regulators about their machine learning model, but they are often not fully certain about their machine learning model's ability to pass the regulations. This also has an impact on their strategic behaviour. Therefore, we model the designer as an strategic agent who is epistemically uncertain about their type i.e. the distribution of their evidence $P(Z)$ with a second order distribution $\theta \in \Delta(\mathcal{P})$ on the space of all possible evidence distributions $\mathcal{P}$. This second order distribution θ represents the prior knowledge of the agent. The model designers want to select a contract from the menu Π such that it maximises their expected utility ex-ante, i.e.

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_\theta[\mathbb{E}_{Z\sim P}[\pi(Z)]] - C \tag{1}$$

Proposition 3.8. *For a menu of contracts Π that enforce obedience to regulation according to definition 3.2, the agent's best response with a second order distribution $\theta \in \Delta(\mathcal{P})$ is to minimise the KL divergence in the convex set of null distributions conv$(\mathcal{P}_0)$ with respect expected distribution $\mathbb{Q}$, i.e.*

$$\pi^* = C\frac{d\mathbb{Q}}{d\mathbb{P}^*} \quad \textit{where} \quad \mathbb{P}^* = \operatorname*{arg\,min}_{P \in conv(\mathcal{P}_0)} KL(\mathbb{Q}|P)$$

where $\mathbb{Q} = \int_{\mathcal{P}} P\theta(P)dP$ is the marginal density obtained by mariginalising over θ.

What if Model Designer's Cost are Private? Let's denote the maximal revenue an agent could earn if they gain market access from the regulator by R. Naturally $R > C$, i.e. the max revenue for an agent must be more than the flat market entry fees C. However, often in real world agents incur costs before they submit a model for regulatory approval. These could include development costs, operational costs and other expenses which are dependent on multiple factors such as size of the model designer companies and their geographical location. We assume this as the private cost of the model designers. For a model designer with type $P \in \mathcal{P}$, We will index the agents with the distribution their model f produces on the evidence i.e. $P(Z)$ assuming that the distribution inherently determines their type. Let $C(P)$ be model designer's private cost to train their model f. Additionally we assume that there exists a threshold of $C_{max} < R$ which is maximum investment any agent would be willing to make for revenue R. Therefore model designer's willingness to pay to fee is $C_{max} - C(P)$. We run a Vickery auction [Vickrey, 1961] for entry into the market where for N total bidders, an auction to access for $K < N$ license purchase is run. We denote the allocation rule with $X : b \rightarrow \{0, 1\}$ denoting which bidder is given access to purchase or not based on their bids. Based on bids, we sort the agents i.e. for $B = \{b(P) \quad \forall P \in \mathcal{P}\}$ we denote the sorted bids as sort(B). Selected designers pay price $r :=$ sort$(B)_{k+1}$ which is the $k + 1^{th}$ highest bid overall. This ensures that the truthful reporting of each agents willingness to pay $b(P) = C_{max} - C(P)$ is the dominant strategy for every model designer and that the payment scheme is uniquely implements [Myerson, 1981]. All designers pay a fees of r allowing us to discriminate between them based on their bids i.e.

$$\sup_{\pi\in\Pi} \mathbb{E}_P[\pi_P(z)] \leq C_{max} - b(P) + r \quad \forall \quad P \in \mathcal{P}_0$$
$$\sup_{\pi\in\Pi} \mathbb{E}_P[\pi_P(z)] \leq C_{max} - (C_{max} - C(P)) + r \quad \forall \quad P \in \mathcal{P}_0$$
$$\sup_{\pi\in\Pi} \mathbb{E}_P[\pi_P(z)] \leq C(P) + r \quad \forall \quad P \in \mathcal{P}_0$$

Inherently, there is a tradeoff for the model designers, investing more by making $C(P)$ larger makes them lose the bidding war as their true valuation for willingness to pay decreases, thus decreasing the chances to license access but investing more in computing improves their later payoff.

3.2 Connection to Classical Hypothesis Testing

We now formulate our regulation problem as a classical testing procedure in a non-parametrized fashion to highlight its differences from incentive aware statistical protocol introduced above. Our regulation can be formulated statistical decision making problem via a composite hypothesis test as follows:

$$H_0: \quad P \in \mathcal{P}_0 \quad H_1: \quad P \in \mathcal{P}_1$$

where $\mathcal{P}_0$ and $\mathcal{P}_1$ are simply the composite null and alternative risk distributions defined as $\mathcal{P}_0 := \{P \mid \mathbb{E}_{Z\sim P}[Z] > \epsilon\}$ and $\mathcal{P}_1 := \{P \mid \mathbb{E}_{Z\sim P}[Z] \leq \epsilon\}$ in the case of ϵ risk control. Notice that the set of parameters $\mathcal{P}_0$ and $\mathcal{P}_1$ characterise the parameters of the risk distributions arising from ϵ safe and un-safe models according to the Definition 3.1. While classical hypothesis tests are a valid protocol to ensure regulation they do not incorporate the incentives of the model designers into the problem setup. With the knowledge of the false positive rate α, incentive aware model designers can then be strategic to include just enough representation into their training data in order to appear safe if the false positive rate α is large enough or the gains under α are large enough to justify their cost-benefit calculus. Thus the classic tests set up binary incentives for the model designers making them incentive incompatible [Bates et al., 2022, Hossain et al., 2025]. The model designer's optimal response introduced in Proposition 3.8, π^* indeed resembles likelihood ratio and for case where $\mathcal{P}_0 = \{P\}$, the best response is indeed likelihood ratio scaled by fees C. In the testing literature, it is very well known that the optimal test for singleton nulls and alternates are likelihood ratio tests [Neyman and Pearson, 1933]. However, one must note that although related the best response π^* is not the same as optimal test in all scenarios since π^* is based on gambles optimise for the growth of revenue (wealth of model provider) and a optimal test optimises for true positives in decision-making [Ramdas and Wang, 2024].

4 Simulations

Hypothesis test for effective model dimension. We consider a toy linear model for our simulations. We further assume that it aims to approximate an original data generating process of $y_i = x_i^T\theta^* + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma)$. Additionally, assume that all features affect the model prediction equally. We now wish to regulate the number of parameters / features used by this model. Either the model designer could be using all the $d = d_0 + 1$ features to make the prediction, i.e., designer also uses the sensitive attribute to maximise their prediction capability or the designer could follow the regulation and only use non-sensitive attributes to make prediction. We frame the use of sensitive attribute as the null hypothesis and use of only non sensitive attribute as alternative to build an hypothesis test for regulation. We test the null hypothesis against the alternative as follows

$$H_0: \text{ Model is using sensitive attribute } d = d_0 + 1,$$
$$H_1: \text{ Model is not using sensitive attribute } d = d_0.$$

We consider the standardized quadratic error for features X under OLS estimator $\hat{\theta}$ as the test statistic. That is $Q_{\text{std}} = \frac{n}{\sigma^2}Q = \frac{n}{\sigma^2}(\hat{\theta} - \theta^*)^\top(XX^\top)(\hat{\theta} - \theta^*)$ which is a χ^2_d distributed and which under suitable regularity conditions follows a chi-square distribution with degrees of freedom equal to the effective number of parameters used in the model (See Appendix A.4 for derivation). Since we know the parametric model of both null and alternative distributions and they are simple singleton hypothesis we can use a likelihood ratio test as it is the optimal test given Neyman Pearson Lemma. Let us denote the test statistic $L := \frac{L(d_0+1;Q)}{L(d_0;Q)}$ where $Q \sim \chi^2_d$ and $L(d;Q)$ is the likelihood of sample Q from chi-squared distribution with parameter d. Under H_0, assuming that the test statistic has a distribution $P_{H_0}(L)$. We implement the test to reject H_0 if $L > \tau_\alpha$ where τ_α is the $1-\alpha$ quantile of $P_{H_0}(L)$, so that we obtain strict α type 1 error. Figure 1 shows that for reasonable $n = 80$ the test has power 1 under type I error control. Under these ideal testing conditions it becomes easy for us to now illustrate the strategic aspects of enforcing regulations via hypothesis tests.

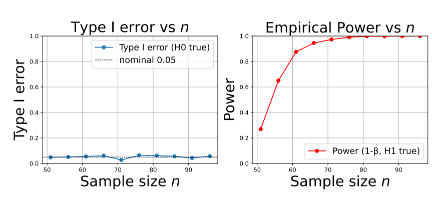
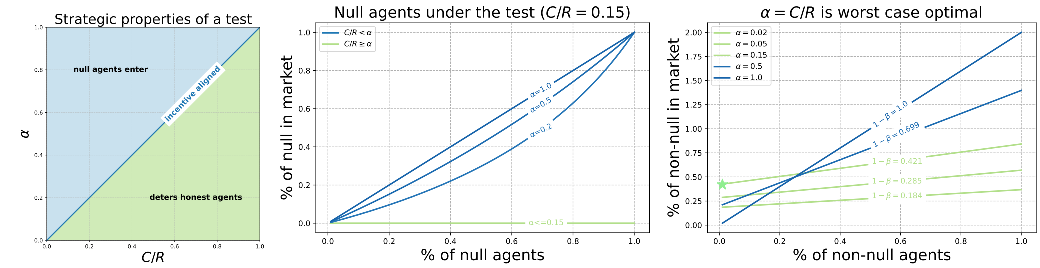


Figure 1: Power vs Type 1 in the Chi-squared test of model parameters/features i.e. $d = 50$ and another sensitive attribute

Strategic Aspects in the test The above testing procedure for enforcing regulation ignores the incentives to the model designers. However, in real world the model designers operate under incentives. In this section we consider some incentives that designers may have and try to understand their behaviour under a statistical test for regulation of use of sensitive attributes for training. Let us assume that for regulation tests the regulator charges a fee, this can also be understood as the tax to operate in the market, we denote it using C and we assume that the size of the market is denoted by R. Ideally a regulation implemented by a test must deter null agents from entering the market, i.e. non obedient agents self opt out of the market while the keeping the obedient agents in the market. With the statistical test proposed above to check the effective dimension of

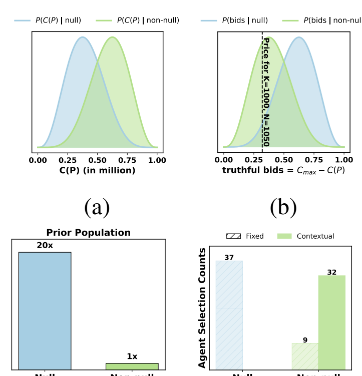


(a) (b) (c)

Figure 2: The strategic reaction of null and non-null agents in the market to regulations via testing. The above figures (b) and (c) assume the incentives in the market by fixing $C/R = 0.15$

the model and thus for the use of sensitive attribute, let us assume that the final test implemented by the regulator has false positive rate α, the type II error $\beta(\alpha)$ is denoted as a function of the choice α made by the regulator thus the choice of false positive rate also dictates the power of the test $1 - \beta(\alpha)$. In an Ex-ante analysis we can observe that null agents participation in the market depends directly on this α, as for a null agent, $\alpha R \geq C$ means that the gamble to enter into the market has net positive expected utility. Thus for $\alpha > \frac{C}{R}$, the market will see full participation for approval by the null agents (see Fig 2 a), and because of test properties α proportion of the nul agents will also get approved (see Fig 2 b). Whereas for the non-null agents, the decision to participate in the market depends upon the power of the test i.e. $(1 - \beta(\alpha))R \geq C$ which can be seen in the Fig 2 c that the too strict value of α results in a power below 0.15 and thus lower than $\frac{C}{R}$ resulting in no participants in the market. As α gradually increases to $\alpha = 0.15$ and thus equal to $\frac{C}{R}$ the power of the test increases resulting in more and more non-null agents being approved and eventually for further increase in α, null agents also start to participate in market.

Strategic Behaviour with private costs In Figure 3 we aim to show that incentive aware regulations when ignoring the diversity in development costs, fail to achieve their purpose and therefore must account for the private costs of the agents. We assume total $N = 1050$ model service providers and Figure 3(c) shows the prior distribution of null and non-null agents, and assume that there are 20x more null agents than non-null agents. This results in 1000 null agents and 50 non-null agents naturally. We also assume private costs for each agent shown in million in Figure 3(a). Naturally, on average the non-null agents have to invest more in satisfying the requirements while null agents on average would have a lower cost. This reflects also in the bids of the agents (Figure 3(b)) where null agents are willing to bid more to gain access to the market because their costs are less, we assume $C_{max} = 1$ million when plotting the bid distributions of null and non null agents. We run a Vickery auction for selecting $K = 1000$ top bidders who pay the 1001^{th} bid as their price for market entry. This allows us to identify the private costs of the providers and now gambles are sold to them contextually. In Figure 3(d) we



(a) (b) (c) (d)

Figure 3: The strategic reaction of null and non-null agents to regulations under private costs.

compare this to setup where regulator assumes a flat fees and evaluates incentive aligned $C/R = 0.05$, for a market where $R = 10$ million and the regulator sets $C = 500$k as the flat fee. Since the flat fee ignores the private costs, it benefits the null agents and hurts the non-null agents, forfeiting the purpose of deterring null agents from entering the contract. Due the cost distribution of null agents, for total of 740 out of 1000 null agents, the flat fee makes it desirable for them to enter the gamble. Therefore they apply to enter the market and C/R-percent of them get through the test. Whereas in case of regulation which is private cost aware, null agents are strategically deterred from entering the market. Since no agent is willing to pay more than $C_{max} = 1$ million to enter the market setting a flat fee of $C = 500$k means only the non-null agents whose private cost is less 500k would enter the market. As compared to scenario of eliciting private costs where fees is 320k allowing more non-null agents to participate in the market while allowing no null agents to participate in the market. The choice of auction parameter K dictates the number of non-null agents but for all choices of K, with truthful elicitation via auction, null agents do not enter the market. In general larger values of K as a hyper-parameter allow for larger market, however, in K is too large then the bidders may not fear competition allowing them to strategically lie about their bids, anticipating access almost always.

5 Related Works

Strategic aspects in Hypothesis Testing Recently there has been a large interest in intersection of economics and hypothesis testing [Bates et al., 2022, 2023, Min, 2023, Hossain et al., 2025]. Applications to clinical trails have motivated study of stragetic aspects within classical hypothesis tests, where Bates et al. [2022, 2023], Min [2023] follow a principal agent framework for their strategic hypothesis test assuming that the principal knows nothing about the distribution of the types of agents, while Hossain et al. [2025] follow a Bayesian game theoretic framework by assuming the distribution of types of agents. Our work is similar to that of the Bates et al. [2022, 2023], Min [2023] in the spirit that we model the problem of regulating model providers also as an principal agent problem in clinical trials where the FDA acts as a regulator and the drug designers as the agents. However, our setup argues for an analysis where the cost of developement of the model is also additonal private information that the model providers possess and the license menus offered to the agents depends on it. Additionally, in our setup of model regulation it is unrealistic to assume that the provider exactly knows their type, i.e. the model they train obeys the regulation or not, which is relaxed by assuming a second order distribution which acts as prior knowledge of the agent that reflects their private information.

Game Theoretic Probability and Auditing While our results characterises the menu of regulation enforcing licenses as desirable gambles leveraging tools from Imprecise Probability [Walley, 1991, Augustin et al., 2014, Crane, 2018] and theory of desirabililty [De Cooman and Quaeghebeur, 2012, De Bock, 2023] an alternative characterisation of regulation enforcing licenses is also possible via e-variables [Shafer, 2021, Vovk and Shafer, Ramdas et al., 2023, Grünwald et al., 2024] as provided in Bates et al. [2022]. E-values or variables have become the standard back bone of methods that perform auditing of machine learning models via sequential tests with applications in risk monitoring and control [Bates et al., 2021, Waudby-Smith and Ramdas, 2024, Timans et al., 2025], fairness [Chugg et al., 2023], differential privacy [González et al., 2025] and many other applications [Shekhar and Ramdas, 2023, Xu and Ramdas, 2024]. However, unlike our method the sequential testing methods that leverage e-variables rather focus on anytime valid statistical inference and do not account from incentives of strategic agents, keeping their betting interpretation only as a didactic tool to demonstrate the validity of their tests.

AI governance and Regulation The rapid development in AI with recent breakthroughs has allowed deployment of AI in the real world. Which consequentially have its societal impacts, causing significant interest in the economists, political scientists, law and policy makers on topics related to AI governance and regulation of AI. This has resulted in a spark of research on topics related to AI governance and regulation that approaches theses questions from ethical [Jobin et al., 2019, Hagendorff, 2020, Huang et al., 2022], policy [Diakopoulos, 2016, O'neil, 2017], socio-cultural [Awad et al., 2018, Vesnic-Alujevic et al., 2020] and political [Pavel et al., 2023, Schmid et al., 2025] perspectives. With our research we aim to surface some technical challenges in achieving the normative requirements that the policy makers aim for and also provide new insights into the operationalising these normative requirements.

6 Conclusion

Statistical contracts embed mechanism design into rulemaking, turning noisy, sample-based verification into an incentive-compatible enforcement mechanism that maps observed model performance to licensed market outcomes and thereby induces non-compliant providers to self-select out while keeping compliant providers viable. This treatment of enforcement as an economic design problem offers a new perspective that can circumvent several prior challenges in AI regulation by shifting the regulator's burden from perfect verification to carefully designed incentives. Allowing regulators to refrain from costly, exhaustive monitoring toward economically sustainable rules that exploit private incentives and sampling noise rather than being defeated by it. The approach reduces the need for perfect verification, limits opportunities for benchmark gaming, and clarifies tradeoffs between monitoring cost, statistical power, and market participation. It is not a panacea, the contracts must be carefully calibrated and paired with provenance and cross-jurisdictional governance. However, it offers a practical, theory-grounded path for regulators to enforce trustworthy AI in environments where traditional verification is infeasible.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Matthias Aßenmacher, and Christoph Jansen. Statistical multicriteria evaluation of llm-generated text. *arXiv preprint arXiv:2506.18082*, 2025.

Thomas Augustin, Frank P. A. Coolen, Gert De Cooman, and Matthias C. M. Troffaes, editors. *Introduction to imprecise probabilities*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2014. ISBN 978-0-470-97381-3.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635, 2003.

Dean W Ball. A framework for the private governance of frontier artificial intelligence. *arXiv preprint arXiv:2504.11501*, 2025.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Stephen Bates, Michael I Jordan, Michael Sklar, and Jake A Soloff. Principal-agent hypothesis testing. *arXiv preprint arXiv:2205.06812*, 2022.

Stephen Bates, Michael I Jordan, Michael Sklar, and Jake A Soloff. Incentive-theoretic bayesian inference for collaborative science. *arXiv preprint arXiv:2307.03748*, 2023.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Andrea Bertolini and Francesca Episcopo. The expert group's report on liability for artificial intelligence and other emerging digital technologies: a critical assessment. *European Journal of Risk Regulation*, 12(3):644–659, 2021.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pages 1006–1014. PMLR, 2015.

Miles Brundage, Shahar Avin, Jack Clark, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Technical report, 2018. Accessed 2025-10-15.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Ben Chugg, Santiago Cortes-Gomez, Bryan Wilder, and Aaditya Ramdas. Auditing fairness by betting. *Advances in Neural Information Processing Systems*, 36:6070–6091, 2023.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

HARRY Crane. The fundamental principle of probability: Resolving the replication crisis with skin in the game. *Researchers. One, under review www. researchers. one/article/2018-08-16*, 2018.

Patricia Gomes Rêgo De Almeida, Carlos Denner Dos Santos, and Josivania Silva Farias. Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology*, 23(3): 505–525, 2021.

Jasper De Bock. A theory of desirable things. In *International Symposium on Imprecise Probability: Theories and Applications*, pages 141–152. PMLR, 2023.

Gert De Cooman and Erik Quaeghebeur. Exchangeability and sets of desirable gambles. *International Journal of Approximate Reasoning*, 53(3):363–395, 2012.

Bruno de Finetti. *Theory of Probability*. John Wiley & Sons, 1974.

Rabanus Derr and Robert C Williamson. Four facets of forecast felicity: Calibration, predictiveness, randomness and regret. *arXiv preprint arXiv:2401.14483*, 2024.

Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.

Cian Eastwood, Alexander Robey, and Shashank Singh. Probable Domain Generalization via Quantile Risk Minimization.

Lilian Edwards. The eu ai act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*, 1:25, 2021.

Tomás González, Mateo Dulce-Rubio, Aaditya Ramdas, and Mónica Ribero. Sequentially auditing differential privacy. *arXiv preprint arXiv:2509.07055*, 2025.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024.

Gillian K Hadfield and Jack Clark. Regulatory markets: The future of ai governance. *arXiv preprint arXiv:2304.04914*, 2023.

Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and machines*, 30(1): 99–120, 2020.

Moritz Hardt. The emerging science of machine learning benchmarks. Online at `https://mlbenchmarks.org`, 2025. Manuscript.

Safwan Hossain, Yatong Chen, and Yiling Chen. Strategic hypothesis testing. *arXiv preprint arXiv:2508.03289*, 2025.

Changwu Huang, Zeqi Zhang, Bifei Mao, and Xin Yao. An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819, 2022.

Christoph Jansen, Georg Schollmeyer, Julian Rodemann, Hannah Blocher, and Thomas Augustin. Statistical multicriteria benchmarking via the gsd-front. *Advances in Neural Information Processing Systems*, 37:98143–98179, 2024.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.

Anton Korinek and Jai Vipra. Concentrating intelligence: scaling and market structure in artificial intelligence. *Economic Policy*, 40(121):225–256, 2025.

Joshua Alexander Kroll. *Accountable algorithms*. PhD thesis, Princeton University, 2015.

Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32, 2024.

Yueqi Li and Sanjay Goel. Making it possible for the auditing of ai: A systematic review of ai audits and ai auditability. *Information Systems Frontiers*, 27(3):1121–1151, 2025.

A Lohn and M Musser. How much longer can computing power, drive artificial intelligence progress?, 2022.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Daehong Min. Screening for experiments. *Games and Economic Behavior*, 142:73–100, 2023.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

Barry Pavel, Ivana Ke, Michael Spirtas, James Ryseff, Lea Sabbag, Gregory Smith, Keller Scholl, and Domenique Lumpkin. Ai and geopolitics: How might ai affect the rise and fall of nations? 2023.

Inioluwa Deborah Raji et al. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020. Accessed 2025-10-15.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *arXiv preprint arXiv:2410.23614*, 2024.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization, April 2020. URL `http://arxiv.org/abs/1911.08731`. arXiv:1911.08731 [cs, stat].

Stefka Schmid, Daniel Lambach, Carlo Diehl, and Christian Reuter. Arms race or innovation race? geopolitical ai development. *Geopolitics*, pages 1–30, 2025.

Emmanouil Seferis, Simon Burton, and Stefanos Kollias. Randomized smoothing (almost) in real time? In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431, 2021.

Shubhanshu Shekhar and Aaditya Ramdas. Reducing sequential change detection to sequential estimation. *arXiv preprint arXiv:2309.09111*, 2023.

Anurag Singh, Mahalakshmi Sabanayagam, Krikamol Muandet, and Debarghya Ghoshdastidar. Robust feature inference: A test-time defense strategy using spectral projections. *arXiv preprint arXiv:2307.11672*, 2023.

Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. In *International conference on machine learning*, pages 5389–5400. PMLR, 2024.

Merlin Stein, Milan Gandhi, Theresa Kriecherbauer, Amin Oueslati, and Robert Trager. Public vs private bodies: Who should run advanced ai evaluations and audits? a three-step logic based on case studies of high-risk industries. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1401–1415, 2024.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). *NIST*, 2023.

Alexander Timans, Rajeev Verma, Eric Nalisnick, and Christian A Naesseth. On continuous monitoring of risk violations under unknown shift. *arXiv preprint arXiv:2506.16416*, 2025.

Philip Moreira Tomei, Rupal Jain, and Matija Franklin. Ai governance through markets. *arXiv preprint arXiv:2501.17755*, 2025.

Matthias C. M. Troffaes and Gert de Cooman. Lower previsions. In *Introduction to Imprecise Probabilities*, pages 159–181. John Wiley & Sons, Chichester, 2014.

Matthias CM Troffaes. Decision making under uncertainty using imprecise probabilities. *International journal of approximate reasoning*, 45(1):17–29, 2007.

UK AI Safety Summit. Capabilities and risks from frontier ai. Technical report, 2023. Accessed 2025-10-15.

Unesco. *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, 2021.

Lucia Vesnic-Alujevic, Susana Nascimento, and Alexandre Polvora. Societal and ethical impacts of artificial intelligence: Critical notes on european policy frameworks. *Telecommunications Policy*, 44(6):101961, 2020.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Vladimir Vovk and Glenn Shafer. Game-theoretic probability. *Introduction to Imprecise Probabilities*.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *International Conference on Artificial Intelligence and Statistics*, pages 3997–4005. PMLR, 2024.

A Appendix

A.1 Proof of Proposition

($\Rightarrow$)

Let us assume that the set of contracts Π satisfy Obedience to regulation (Def 3.2) i.e.

$$\begin{aligned}
&\sup_{\pi\in\Pi} \mathbb{E}_{Z\sim P}[\pi(Z)] \leq C \quad \forall P \in \mathcal{P}_0 \\
\implies \quad &\sup_{\pi\in\Pi}\ \sup_{P\in\mathcal{P}_0}\ \mathbb{E}_{Z\sim P}[\pi(Z)] \leq C \\
\implies \quad &\sup_{P\in\mathcal{P}_0}\ \mathbb{E}_{Z\sim P}[\pi(Z) - C] \leq 0 \quad \forall \pi \in \Pi \\
&\inf_{P\in\mathcal{P}_0}\ \mathbb{E}_{Z\sim P}[C - \pi(Z)] \geq 0 \quad \forall \pi \in \Pi
\end{aligned}$$

Which shows that $C - \Pi \subseteq \mathfrak{G}_{\geq 0,\mathcal{P}_0}$.

($\Leftarrow$) Let us assume that $C - \Pi \subseteq \mathfrak{G}_{\geq 0,\mathcal{P}_0}$ i.e. the gambles $C - \Pi$, induced by the license menu are desirable to the Regulator.

$$\begin{aligned}
&\inf_{P\in\mathcal{P}_0}\ \mathbb{E}_{Z\sim P}[C - \pi(Z)] \geq 0 \quad \forall \pi \in \Pi \\
\implies \quad &\sup_{P\in\mathcal{P}_0}\ \mathbb{E}_{Z\sim P}[\pi(Z)] \leq C \quad \forall \pi \in \Pi \\
\implies \quad &\sup_{\pi\in\Pi}\ \sup_{P\in\mathcal{P}_0}\ \mathbb{E}_{Z\sim P}[\pi(Z)] \leq C \\
&\sup_{P\in\mathcal{P}_0}\ \sup_{\pi\in\Pi} \mathbb{E}_{Z\sim P}[\pi(Z)] \leq C \\
&\sup_{\pi\in\Pi} \mathbb{E}_{Z\sim P}[\pi(Z)] \leq C \quad \forall P \in \mathcal{P}_0
\end{aligned}$$

which shows that Π satisfies Definition 3.2.

A.2 Proof of Proposition

Without any loss of generality let us assume that the preference relation $\succ$ on Z is such that smaller the evidence the better, as in case of loss values. Therefore, in order to show that menu Π must be monotone in the preference order of $\mathbb{Z}$, we need to show that every $\pi \in \Pi$ is monotonically decreasing in $\mathbb{Z}$. ($\Rightarrow$) Let us assume that the license function menu Π is monotonically decreasing in z for every $\pi \in \Pi$. That is,

$$\forall z_1, z_2 \in \mathbb{Z} \quad z_1 < z_2 \quad \implies \pi(z_1) > \pi(z_2) \quad \forall \pi \in \Pi$$

From which we can say that, for any two distributions P_1 and P_2 in $\Delta(\mathbb{Z})$, which are induced by models f_1 and f_2, the following holds,

$$\mathbb{E}_{Z\sim P_1}[Z] < \mathbb{E}_{Z\sim P_2}[Z] \quad \implies E_{Z\sim P_1}[\pi(Z)] > E_{Z\sim P_2}[\pi(Z)] \quad \forall \pi \in \Pi \tag{2}$$

Intutively, for reducting the expected value of random variable Z, the distribution P_1 must put larger mass on smaller values of Z as compared to P_2 and since π is monotonically decreasing, for smaller values of Z it will be larger, thus making the expectation $E_{Z\sim P_1}[\pi(Z)]$ larger in comparison to $E_{Z\sim P_2}[\pi(Z)]$.

$$\mathbb{E}_{Z\sim P_1}[Z] < \mathbb{E}_{Z\sim P_2}[Z] \quad \max_{\pi\in\Pi} E_{Z\sim P_1}[\pi(Z)] > \max_{\pi\in\Pi} E_{Z\sim P_2}[\pi(Z)] \quad \text{(from Eq. 2)} \tag{3}$$

Since Equation 2 is elementwise valid for all $\pi \in \Pi$ we can say that $\max_{\pi\in\Pi} E_{Z\sim P_1}[\pi(Z)] > \max_{\pi\in\Pi} E_{Z\sim P_2}[\pi(Z)]$ and thus Π is incentive compatible.

A.3 Proof of Proposition

$$\pi^* = \arg\max_{\pi\in\Pi} \mathbb{E}_\theta[\mathbb{E}_{Z\sim P}[\pi(Z)]] - C$$

Then the above optimisation task can be re-written as

$$\begin{aligned}
&\max_{\pi:Z\to\mathbb{R}_{\geq 0}} \mathbb{E}_\theta[\mathbb{E}_{Z\sim P}[\pi(Z)]] \quad \text{subject to } \sup_{P\in\mathcal{P}_0} \mathbb{E}_P[\pi(Z)] \leq C \\
&\max_{\pi:Z\to\mathbb{R}_{\geq 0}} \int_{P\in\mathcal{P}} \left[\int_Z \pi(Z)P(Z)dZ\right] d\theta(P) \quad \text{subject to } \sup_{P\in\mathcal{P}_0} \mathbb{E}_P[\pi(Z)] \leq C \\
&\max_{\pi:Z\to\mathbb{R}_{\geq 0}} \int_Z \pi(Z)\left[\int_{P\in\mathcal{P}} P(Z)d\theta(P)\right] dZ \quad \text{subject to } \sup_{P\in\mathcal{P}_0} \mathbb{E}_P[\pi(Z)] \leq C \\
&\max_{\pi:Z\to\mathbb{R}_{\geq 0}} \mathbb{E}_{Z\sim\int P(Z)d\theta(P)}[\pi(Z)] \quad \text{subject to } \sup_{P\in\mathcal{P}_0} \mathbb{E}_P[\pi(Z)] \leq C \\
&\max_{\pi:Z\to\mathbb{R}_{\geq 0}} \mathbb{E}_{Z\sim\mathbb{Q}}[\pi(Z)] \quad \text{subject to } \sup_{P\in\mathcal{P}_0} \mathbb{E}_P[\pi(Z)] \leq C
\end{aligned}$$

where $\mathbb{Q}$ is the mariginalised distribution over θ. Since θ was a second order distribution Q is a valid distribution over Z. Also notice that then the above problem translates to a known problem of testing a simple point alternative against a composite null $\mathcal{P}_0$ once we scale $\varphi := \frac{1}{C}\pi$ then the above problem is equivalent to,

$$\max_{\varphi:Z\to[0,1]} \mathbb{E}_{Z\sim\mathbb{Q}}[\varphi(Z)] \quad \text{subject to } \sup_{P\in\mathcal{P}_0} \mathbb{E}_P[\varphi(Z)] \leq 1$$

This problem has an optimal solution via Reverse Information Projection (RiPr) which is defined as the numeraire e-variable and also shown to be unique (See Chapter 6 [Ramdas and Wang, 2024]). Formally,

$$\phi^*(Z) = \frac{d\mathbb{Q}}{d\mathbb{P}^*} \quad \text{where} \quad \mathbb{P}^* = \arg\min_{P\in\text{conv}(\mathcal{P}_0)} KL(P||\mathbb{Q})$$

And therefore $\pi^* = C\frac{d\mathbb{Q}}{d\mathbb{P}^*}$.

A.4 Distribution of Excess Risk

We consider a fixed-design linear model with Gaussian noise. The data-generating process is

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}^* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

where $y_i \in \mathbb{R}$ is the observed output, $\boldsymbol{x}_i \in \mathbb{R}^d$ is the input and $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is the true (unknown) parameter vector, $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ is the fixed design matrix. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ is a zero-mean Gaussian noise vector with independent entries and variance σ^2. Thus in the matrix notation we will write $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}$. The agent observes $(\boldsymbol{X}, \boldsymbol{y})$ and estimates $\boldsymbol{\theta}^*$ via Ordinary Least Squares (OLS):

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{\theta}^* + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}. \tag{5}$$

Under these assumptions, the covariance of the estimator is $\mathrm{Cov}(\hat{\theta}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \frac{\sigma^2}{n} \hat{\Sigma}^{-1}$.

The agent's risk for some parameter θ, evaluated using a positive semi-definite matrix $\hat{\Sigma} = \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X}$, is

$$\begin{aligned}
R(\theta) &= \frac{1}{n} \mathbb{E}_{\boldsymbol{y}}[||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2] = \frac{1}{n} \mathbb{E}_{\boldsymbol{\epsilon}}[||\boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon} - \boldsymbol{X}\boldsymbol{\theta}||_2^2] \\
&= \frac{1}{n} \mathbb{E}_{\boldsymbol{\epsilon}}[||\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)||_2^2 + \boldsymbol{\epsilon}^T (\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}^*) + ||\boldsymbol{\epsilon}||_2^2] \\
&= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \hat{\Sigma} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \sigma^2 \\
&= ||\boldsymbol{\theta} - \boldsymbol{\theta}^*||_{\hat{\Sigma}} + \sigma^2
\end{aligned}$$

.

The expected risk is then

$$\begin{aligned}
\mathbb{E}[R(\hat{\theta})] &= \mathbb{E}[||\boldsymbol{\theta} - \boldsymbol{\theta}^*||_{\hat{\Sigma}}] + \sigma^2 \\
&= \frac{1}{n} \mathbb{E}[\epsilon \top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \epsilon] + \sigma^2 \\
&= \frac{1}{n} \mathbb{E}[\epsilon \top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \epsilon] + \sigma^2 \\
&= \frac{1}{n} \mathbb{E}[tr(\epsilon \top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \epsilon)] + +\sigma^2 \\
&= \frac{1}{n} tr(\mathbb{E}[\epsilon \epsilon^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top]) + \sigma^2 \\
&= \frac{1}{n} tr(\mathbb{E}[\epsilon \epsilon^\top] \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top) + \sigma^2 \\
&= \frac{\sigma^2}{n} \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}) + \sigma^2 = \frac{\sigma^2}{n} \operatorname{tr}(I_d) + \sigma^2 = \frac{\sigma^2 d}{n} + \sigma^2.
\end{aligned}$$

Let us define the excess risk in the quadratic form as

$$Q := R(\hat{\theta}) - \sigma^2 = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \hat{\Sigma} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \tag{6}$$

Since $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{\theta}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\theta}^* + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \epsilon - \boldsymbol{\theta}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \epsilon$ is a linear function of a multivariate Gaussian ϵ, we have

$$\hat{\theta} - \theta^* \sim \mathcal{N}\big(0, \Sigma_{\hat{\theta}}\big), \quad \Sigma_{\hat{\theta}} := \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \frac{\sigma^2}{n} \hat{\Sigma}^{-1}. \tag{7}$$

The excess risk is a quadratic form of a Gaussian:

$$Q = (\hat{\theta} - \theta^*)^\top \Sigma (\hat{\theta} - \theta^*) \sim \text{Chi-squared}. \tag{8}$$

Let ${\Sigma_{\hat{\theta}}}^{1/2}$ denote a square root of $\Sigma_{\hat{\theta}}$. Then

$$Q = (\Sigma_\theta^{1/2} z)^\top \hat{\Sigma} (\Sigma_\theta^{1/2} z), \quad z \sim \mathcal{N}(0, I_d) \tag{9}$$

$$= \frac{\sigma^2}{n} z^\top (\hat{\Sigma}^{-1/2} \hat{\Sigma} \hat{\Sigma}^{-1/2}) z \tag{10}$$

$$=: \frac{\sigma^2}{n} z^\top I_d z \tag{11}$$

$$= \frac{\sigma^2}{n} z^\top z \tag{12}$$

Therefore $Q \sim \frac{\sigma^2}{n}\chi_d$ where χ_d is a Chi-square distribution with degree of freedom d which denote the parameters in our model.

A.5 Challenges beyond statistical issues

Statistical or technical challenges set aside, AI regulation has several non technical challenges in regulation as there is seldom any goods or process that are as general as "intelligence" and have such close human interaction. One key issue is that the liability of AI model's risk is fragmented across model designers, data suppliers, integrators, and deployers, complicating enforcement [Tabassi, 2023, Bertolini and Episcopo, 2021]. Another aspect Jurisdictional fragmentation and cross-border deployment, which undermine coherent remedies and legal actions on designers or other stake holders[Edwards, 2021, UK AI Safety Summit, 2023]. There are no widely adopted technical standards or certification regimes; proprietary intellectual-property and trade secrets conflict with transparency and auditability [Raji et al., 2020]. Supply-chain opacity in data provenance, labeling, and collection prevents reliable forensics also offer some additional challenges[Bender et al., 2021]. Market concentration of some large scale service providers also known as "big-tech" in compute and data creates political-economy pressures and regulatory capture [Lohn and Musser, 2022, Korinek and Vipra, 2025]. Dual-use capabilities, adversarial gaming, and benchmark overfitting lets actors satisfy narrow tests while retaining harmful capacity [Blum and Hardt, 2015, Mazeika et al., 2024, Hardt, 2025]. Often evidence standards in courts and agencies are immature for probabilistic, high-dimensional technical proofs [Kroll, 2015]. Certification and continuous audit impose high fixed costs that raise market-entry barriers [Raji et al., 2020]. Human-in-the-loop requirements are hard to specify and brittle in practice [Amodei et al., 2016]. Finally, cultural and ethical pluralism, privacy tradeoffs in monitoring, and systemic risks from correlated deployments mean regulation must reconcile competing values under uncertainty [Unesco, 2022]. These legal, economic, organizational, and security frictions interact with the two-layer statistical uncertainty to make AI regulation both harder to formulate and easier for stakeholders to plausibly evade than conventional product regulation [Brundage et al., 2018].