ADVANCING REGULATION IN ARTIFICIAL INTELLIGENCE: AN AUCTION-BASED APPROACH

Marco Bornstein[*], **Zora Che, Suhas Julapalli, Abdirisak Mohamed, & Furong Huang**
Department of Computer Science, University of Maryland, College Park, MD, USA

Amrit Singh Bedi
Department of Computer Science, University of Central Florida, Orlando, FL, USA

ABSTRACT

In an era of "moving fast and breaking things", regulators have moved slowly to pick up the safety, bias, and legal debris left in the wake of broken Artificial Intelligence (AI) deployment. While there is much-warranted discussion about how to address the safety, bias, and legal woes of state-of-the-art AI models, rigorous and realistic mathematical frameworks to regulate AI are lacking. Our paper addresses this challenge, proposing an auction-based regulatory mechanism that provably incentivizes devices (i) to deploy compliant models and (ii) to participate in the regulation process. We formulate AI regulation as an all-pay auction where enterprises submit models for approval. The regulator enforces compliance thresholds and further rewards models exhibiting higher compliance than their peers. We derive Nash Equilibria demonstrating that rational agents will submit models exceeding the prescribed compliance threshold. Empirical results show that our regulatory auction boosts compliance rates by 20% and participation rates by 15% compared to baseline regulatory mechanisms, outperforming simpler frameworks that merely impose minimum compliance standards.

1 INTRODUCTION

The recent large-scale deployment of artificial intelligence (AI) models, such as large language models (LLMs), has simultaneously boosted human productivity while sparking concern over safety (*e.g.,* hallucinations, bias, and privacy (Huang et al., 2025)). Many industry leaders, such as Google and OpenAI, remain embroiled in controversy surrounding bias and misinformation (Brewster, 2024; Robertson, 2024; White, 2024), safety (Jacob, 2024; Seetharaman, 2024; White, 2023), as well as legality and ethics (Bruell, 2023; Metz et al., 2024; Moreno, 2023) in their development and deployment of LLMs. Furthermore, irresponsible LLM deployment risks the spread of misinformation or propaganda by adversaries (Barman et al., 2024; Neumann et al., 2024; Sun et al., 2024). To date, a consistent and industry-wide solution to oversee responsible AI deployment remains elusive.

Naturally, one solution to mitigate these dangers is to increase governmental regulation over AI deployment. In the United States, there have been some strides, on federal (House, 2023) and state levels (Information, 2024), to regulate the safety and security of large-scale AI systems (including LLMs). While these recent executive orders and bills highlight the necessity to develop safety standards and enact safety and security protocols, few details are offered. This follows a consistent trend of well-deserved scrutiny towards the lack of AI regulation without providing an answer on *how to develop rigorous and realistic mathematical frameworks to achieve AI regulation*.

We believe that a rigorous and realistic mathematical framework for AI regulation consists of four key pieces: **(a)** the ability to model and to analyze participant decisions, **(b)** the existence of an "optimal" participant equilibrium, **(c)** limited mathematical assumptions, and **(d)** straightforward implementation of the framework by a regulator. This work takes a first step towards unlocking each of these four keys, designing a regulatory framework to not only enforce strict compliance, *e.g.,* safety or ethical compliance, of deployed AI models, but simultaneously to incentivize the production of more compliant AI models.

[*] Correspondence to Marco Bornstein: `marcobornsteinresearch@gmail.com`.

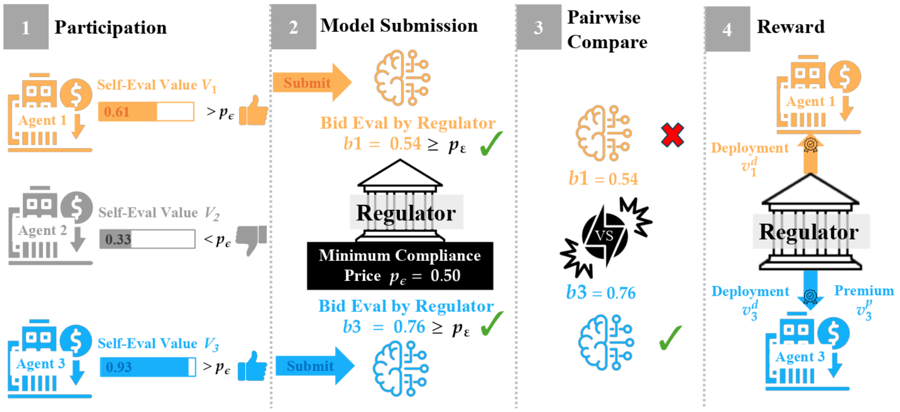


Figure 1: **Step-by-Step CIRCA Schematic.** (Step 0) The regulator sets a compliance threshold, ϵ, having corresponding price, p_ϵ, required to achieve ϵ. (Step 1) Agents evaluate their total value, V_i, from model deployment value (v_i^d) and potential regulator compensation (v_i^p). Agents only participate if their total value exceeds p_ϵ. (Step 2) Participating agents submit their models to the regulator, accompanied by their bid b_i, which reflects the amount spent to improve their model's compliance level. Models with bids below p_ϵ are automatically rejected. (Step 3) The submitted models are randomly paired, and the more compliant model (*i.e.,* the higher bid) in each pair wins. In this example, agent 3 wins since $b_3 > b_1$. (Step 4) Winning models receive both a premium and deployment value (*i.e.,* agent 3 wins premium v_3^p and deployment v_3^d values), while losing models receive only the deployment value (*i.e.,* agent 1 only wins deployment value v_1^d).

Specifically, we **(a)** formulate the AI regulatory process as an *all-pay auction*, where agents (enterprises) submit their models to a regulator. This novel auction-based regulatory mechanism leverages a reward-payment protocol that **(b)** emits Nash Equilibria at which agents *deploy models that are more compliant than a prescribed threshold.* Analysis of our auction-based approach **(c)** requires few assumptions. While inclusion of assumptions is non-ideal, the usage of these assumptions allows us to advance AI regulation within a sparse, yet critical, area of research. We note, however, that the two assumptions used in this work are used within existing regulatory and AI settings (Goulder & Schein, 2013; Howe et al., 2024; Rajpurkar et al., 2016; Stavins, 2008; U.S. Food and Drug Administration, 2022; Williams et al., 2018; Zaremba et al., 2025) (Section 3). Finally, our approach is **(d)** simple and can easily be implemented by a regulator (Figure 1). Like existing regulatory frameworks (Coglianese & Kagan, 2007; Powell, 2014; Van Norman, 2016), we only require the regulator to: (i) prohibit deployment of models that fail to meet prescribed compliance thresholds, and (ii) incentivize compliant model production and deployment by providing additional rewards to agents that submit more compliant models than their peers.

We summarize our contributions as follows:

(1) AI Regulation: We propose a Compliance-Incentivized Regulatory-Centered Auction (CIRCA), offering a novel approach towards AI regulation.

(2) Compliance-First: We establish, through derived Nash Equilibria, that agents are incentivized to submit models surpassing the required compliance threshold.

(3) Effectiveness: Our empirical results show that CIRCA increases model compliance by over 20% and boosts participation rates by 15% compared to baseline regulatory mechanisms.

2 Related Works

Regulation Frameworks for Artificial Intelligence. A handful of work focuses on regulation frameworks for AI deployment (de Almeida et al., 2021; Jagadeesan et al., 2024; Rodríguez et al., 2022). First, de Almeida et al. (2021) details the need for AI regulation and surveys existing proposals. The proposals are ethical frameworks that express ethical decisions to make and dilemmas to address. However, these proposals lack a mathematical framework to incentivize provably compliant models. Rodríguez et al. (2022) utilize AI models to detect collusive auctions. This work is related to our paper but in reverse: Rodríguez et al. (2022) applies AI to regulate auctions and to ensure that they are not collusive. In contrast, our paper aims to use auctions to regulate AI deployment. Jagadeesan et al. (2024) focuses on reducing barriers to entry for smaller companies that are competing against incumbent, larger companies. A multi-objective high-dimensional regression framework is proposed to impose "reputational damage" upon companies that deploy unsafe AI models. This work allows varying levels of safety constraints, where newer companies face less severe constraints in order to spur their entry into the market, which is unrealistic in many settings and only considers simple linear-regression models.

The closest related work to ours, Yaghini et al. (2024), proposes a regulation game for ensuring privacy and fairness that is formulated as a Stackelberg game. This game is a multi-agent optimization problem that is also multi-objective (for fairness and privacy). An equilibrium-search algorithm is presented to ensure that agents remain on the Pareto frontier of their objectives (although this frontier is estimated algorithmically). Notably, Yaghini et al. (2024) considers only one model builder (agent) and multiple regulators that provide updates to the agent's strategy. Here, a more realistic setup is considered, where there are multiple agents and a single regulator whose goal is to incentivize compliant model deployment. It falls out of the scope of a regulator's job to collaborate with agents to optimize their strategy. Furthermore, the mechanism proposed here is simple and efficient. No Pareto frontier estimation or multiple rounds of optimization are required.

All-Pay Auctions. Compared to the dearth of literature in regulatory frameworks for AI, all-pay auctions are well-researched (Amann & Leininger, 1996; Baye et al., 1996; Bhaskar, 2018; DiPalantino & Vojnovic, 2009; Gemp et al., 2022; Goeree & Turner, 2000; Siegel, 2009; Tardos, 2017). These works formulate specific all-pay auctions and determine their equilibria. Some works consider settings where agents have complete information about their rivals' bids (Baye et al., 1996) while others consider incomplete information, such as only knowing the distribution of agent valuations (Amann & Leininger, 1996; Bhaskar, 2018; Tardos, 2017). One major application of all-pay auctions are crowd-sourcing competitions. Many agents participate to win a reward, with those losing incurring a small cost for their time, effort, *etc*. DiPalantino & Vojnovic (2009) is one of the first works to model crowd-sourcing competitions as an all-pay auction. Further research, such as Gemp et al. (2022), have leveraged AI to design all-pay auctions for crowd-sourcing competitions. However, instead of crowd-sourcing, our paper formulates the AI regulatory process as an asymmetric and incomplete all-pay auction. Previous analysis in this setting (Amann & Leininger, 1996; Bhaskar, 2018; Tardos, 2017) is leveraged to derive Nash Equilibria.

3 Regulatory Compliance of Artificial Intelligence

There exists a regulator R with the compliance power to set and to enforce laws and regulations (*e.g.,* U.S. government regulation on lead exposure). The regulator wants to regulate AI model deployment, by ensuring that all models meet a compliance threshold $\epsilon \in (0, 1)$, *e.g.,* the National Institute for Occupational Safety and Health (NIOSH) regulates that N95 respirators filter out at least 95% of airborne particles. If a model does not reach the compliance threshold ϵ, then it is deemed unsafe and the regulator bars deployment. On the other side, there are n rational model-building agents. Agents seek to maximize their own benefit, or utility.

Bidding & Evaluation. By law, each agent i must submit, or bid in auction terminology, its model $w_i \in \mathbb{R}^d$ for evaluation to the regulator before it can be approved for deployment. Let $S(w; x) : \mathbb{R}^d \to \mathbb{R}_+$ output a compliance level (the larger the better) for model w given data x. In effect, each agent, given its own data x_i, bids a compliance level $s_i^A := S(w_i; x_i)$ to the regulator. Subsequently, the regulator, using its own data x_R, independently evaluates the agent's compliance level bid as $s_i^R := S(w_i; x_R)$. Agent and regulator evaluation data is assumed to be independent and identically distributed (IID) $x_i, x_R \sim \mathcal{D}$.

Assumption 1. *Agent and regulator evaluation data comes from the same distribution $x_i, x_R \sim \mathcal{D}$.*

This assumption is realistic, because agents and regulators typically rely on standardized data collection processes (U.S. Food and Drug Administration, 2022) or widely accepted datasets (Rajpurkar et al., 2016; Williams et al., 2018) for evaluation. This ensures a fair and unbiased assessment of compliance. For example, FDA guidelines detail that data collection should assess efficacy and safety across various subgroups, *e.g.,* demographics, while also not changing "baseline data collection determined by the clinical trial objectives" (U.S. Food and Drug Administration, 2022). In areas such as Natural Language Processing, common datasets, or benchmarks, are employed to consistently evaluate model comprehension (Rajpurkar et al., 2016; Williams et al., 2018), trustworthiness (Wang et al., 2023), and security (Munoz et al., 2024) across various models. Therefore, it is reasonable to define agent i's compliance level bid as $s_i := \mathbb{E}_{x\sim\mathcal{D}}[S(w_i; x)]$. The scenario where evaluation data may be non-IID is addressed within Appendix F.

In regulatory settings, like the NIOSH example, a scalar compliance metric is often used. If multiple compliance metrics must be monitored, S can be defined to aggregate and weigh the various metrics. This too is realistic in AI. For example, LLM safety alignment literature uses a scalar-valued reward to ensure a model is aligned (Christiano et al., 2017; Kaufmann et al., 2023; Ouyang et al., 2022).

Price of Compliance. We assume that there exists a strictly increasing function $M : (0, 1) \to (0, 1)$ that determines the "price of compliance" (*i.e.,* maps compliance into cost). Simply put, higher-compliant models cost more to attain. Thus, we define the price of ϵ-compliance as $p_\epsilon := M(\epsilon)$.

Assumption 2. $\epsilon > \epsilon' \implies M(\epsilon) > M(\epsilon')$. *A strictly increasing M maps compliance to cost.*

One prominent existing example of this relationship is the cap-and-trade system that the Environmental Protection Agency exercises to combat pollution (Goulder & Schein, 2013; Stavins, 2008). Companies that pollute above a set emission threshold can reach compliance by purchasing allowances, or pollution deficits, from other compliant companies. Thus, pollution compliance is attained with greater cost. For an example within AI, models can achieve higher safety compliance through Machine Unlearning (Liu et al., 2024) or AI Alignment (Dai et al., 2024). However, such methods incur greater computational and data collection costs in exchange for improved compliance. Furthermore, it has been found empirically that larger models and longer inference attain higher levels of compliance in adversarial training, robustness transfer, and defense (Howe et al., 2024; Zaremba et al., 2025). However, larger models and longer inference increase training and inference costs. We validate the compliance-cost relationship empirically in Section 6.

Agent Costs. Realistically for agents, training a compliant model comes with added cost. Consequently, each agent i must decide how much money to *bid*, or spend, b_i to make its model compliant. By Assumption 2, the compliance level of an agent's model will be $s_i = M^{-1}(b_i)$.

Agent Values. *(1) Model deployment value v_i^d.* While it costs more for agents to produce compliant models, they gain value from having their models deployed. Intuitively, this can be viewed as the expected value v_i^d of agent i's model. The valuation for model deployment varies across agents (*e.g.,* Google may value having its model deployed more than Apple). *(2) Premium reward value v_i^p.* Beyond value for model deployment, the regulator can also offer additional, or premium, compensation valued as v_i^p by agents (*e.g.,* tax credits for electric vehicle producers or Fast Track and Priority Review of important drugs by the U.S. Food & Drug Administration). The regulator provides additional compensation to agents whose models exceed the prescribed compliance threshold. However, the value of this compensation varies across agents due to differing internal valuations. It is unrealistic for the regulator to compensate all agents meeting the compliance threshold due to budget constraints. Therefore, additional rewards are limited to a top-performing half of agents surpassing the threshold. This ensures targeted compensation for agents enhancing compliance while maintaining feasibility for the regulator.

Value Distribution. The total value for each agent i is defined as $V_i := v_i^d + v_i^p$, which represents the sum of the deployment value and premium compensation. Although these values may vary widely in practice, $\{V_i\}_{i=1}^n$ is normalized for all n agents to be between 0 and 1 for analytical tractability, allowing a standardized range. Consequently, the price to achieve the compliance threshold ϵ is also normalized to fall within the $(0, 1)$ interval, *i.e.,* $p_\epsilon \in (0, 1)$. The scaling factor $\lambda_i \sim \mathcal{D}_\lambda(0, 1/2)$ dictates the proportion of total value allocated to deployment versus compensation. Therefore, (i) the deployment value is $v_i^d := (1 - \lambda_i)V_i$, and (ii) the premium compensation value is $v_i^p := \lambda_i V_i$. Both V_i and λ_i are private to each agent, though the distributions $\mathcal{D}_V$ and $\mathcal{D}_\lambda$ are known by participants. The maximum allowable factor is set at $\lambda_i = 1/2$, reflecting the realistic constraint that compensation should not exceed deployment value. Although Section 5 primarily considers $\lambda_i \leq 1/2$, theoretical extensions can be made for scenarios where $\lambda_i > 1/2$.

All-Pay Auction Formulation. Overall, agents face a trade-off: producing higher-compliant models garners value, via the regulator, but incurs greater costs. Furthermore, in order to attain the rewards detailed above, agents must submit a model with a compliance level at least as large as ϵ. This problem is formulated as an *asymmetric all-pay auction* with *incomplete information* (Amann & Leininger, 1996; Bhaskar, 2018; Tardos, 2017). An all-pay auction is used since agents incur an unrecoverable cost, training costs, when submitting their model to regulators. The auction is asymmetric with incomplete information since valuations V_i are private and differ for each agent.

Agent Objective. The objective, for each model-building agent i, is to maximize its own utility u_i. Namely, each agent seeks to determine an optimal compliance level to bid to the regulator b_i. However, given the all-pay auction formulation, agents may need to take into account all other agents' bids $\boldsymbol{b}_{-i}$ in order to determine their optimal bid b_i^*,

$$b_i^* := \arg\max_{b_i} u_i(b_i; \boldsymbol{b}_{-i}). \tag{1}$$

A major portion of this paper is constructing an auction-based mechanism, thereby designing the utility of each agent, such that each participating agent maximizes its utility when each bids more than "the price to obtain the minimum compliance threshold", *i.e.,* $b_i^* > p_\epsilon$. To begin, a simple mechanism is provided, already utilized by regulators, that does not accomplish this goal, before detailing the auction-based mechanism CIRCA that provably ensures that $b_i^* > p_\epsilon$ for all agents.

4 Reserve Thresholding: Base Regulation

The simplest method to ensure model compliance is for the regulators to set a reserve price, or minimum compliance level. This mechanism is coined the *multi-winner reserve thresholding auction*, where the regulator awards a deployment reward, v_i^d, to each agent whose model meets or exceeds the compliance threshold ϵ. Within this auction, each agent i's utility is mathematically formulated as,

$$u_i(b_i; \boldsymbol{b}_{-i}) = \begin{cases} -b_i \text{ if } b_i < p_\epsilon, \\ v_i^d - b_i \text{ if } b_i \geq p_\epsilon. \end{cases} \tag{2}$$

However, the formulation above is ineffective at incentivizing greater than ϵ-level compliance.

Theorem 1 (Reserve Thresholding Nash Equilibrium)**.** *Under Assumption 2, agents participating in Reserve Thresholding Equation 2 have an optimal bid and utility of,*

$$b_i^* = p_\epsilon, \quad u_i(b_i^*; \boldsymbol{b}_{-i}) = v_i^d - p_\epsilon, \tag{3}$$

and submit models with the following compliance level,

$$s_i^* = \epsilon \;\; \textit{if} \;\; u_i(b_i^*; \boldsymbol{b}_{-i}) > 0, \quad \textit{else} \;\; s_i^* = 0 \;\; \textit{(no submission).} \tag{4}$$

When a regulator implements reserve thresholding, as formally detailed in Theorem 1, agents exert minimal effort, submitting models that just meet the required compliance threshold ϵ. While this approach ensures that all deployed models satisfy minimum compliance, it fails to encourage agents to build models with compliance levels exceeding ϵ. Additionally, agents whose deployment rewards are less than the cost of achieving compliance, *i.e.,* $v_i^d < p_\epsilon$, lack incentive to participate in the regulatory process. That lack of incentive leads to reduced participation rates (Remark 1).

Remark 1 (Lack of Incentive)**.** *Each agent is only incentivized to submit a model with compliance $s_i^* = \epsilon$. Our goal is to construct a mechanism that incentivizes agents to build models that possess compliance levels exceeding the minimum threshold: $s_i^* > \epsilon$.*

5 Compliance-Incentivized Regulation: Auction-Based Approach

To alleviate the lack of incentives within simple regulatory auctions, such as the one in Section 4, we propose a regulatory all-pay auction that mandates an equilibrium where agents *submit models with compliance levels exceeding* ϵ.

Algorithm Description. The core component of the auction is that agent compliance levels are compared against one another, with the regulator rewarding those having the superior compliant model with premium compensation. Only agents with models that achieve a compliance level of ϵ or higher are eligible to participate in the comparison process; models that do not meet this threshold are automatically rejected. The detailed algorithmic block of CIRCA is depicted in Algorithm 1.

Algorithm 1 Compliance-Incentivized Regulatory-Centered Auction (Circa)

```
Each agent $i$ receives their total value $V_i$ and partition ratio $\lambda_i$ from "nature"
Agents determine their optimal bids $b_i^*$ and corresponding utility $u_i(b_i^*)$ ▷ via Corollaries 1 or 2
Agents decide to participate, the set of participating agents is $P = \{j \in [n] \mid u_j(b_j^*;\ \mathbf{b}_{-j}, v_j^d, v_j^p) > 0\}$
for participating agents $j \in P$ do
    Spend $b_j^*$ to build a model, with compliance $s_j^* = M^{-1}(b_j^*)$, and submit it to the regulator
end for
Regulator verifies compliance levels, clearing models for deployment when $s_j^* \geq \epsilon\ \forall j \in P$
Regulator pairs up models, awarding compensation to agents with the more compliant model
```

Agent Utility. The utility for each agent i is therefore defined as in Equation 5,

$$u_i(b;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = \big(v_i^d + v_i^p \cdot 1_{(b>b_j)}\big) \cdot 1_{(b \geq p_\epsilon)} - b. \tag{5}$$

Per regulation guidelines, the compliance criteria of an accepted model must at least be ϵ. Equation 5 dictates that values are only realized by each agent if its model produces a bid larger than the required cost to reach ϵ-compliance, $1_{(b \geq p_\epsilon)}$. Furthermore, agents only realize additional compensation value v_i^p from the regulator if their compliance level outperforms a randomly selected agent j, $1_{(b>b_j)}$. Any agent that bids $b = 1$ will automatically win and realize both v_i^p and v_j^w. It is important to note that the cost that every agent incurs when building its model is sunk: if the model is not cleared for deployment, the cost $-b$ is still incurred. The agent utility is rewritten in a piece-wise manner below,

$$u_i(b;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = \begin{cases} -b & \text{if } b < p_\epsilon, \\ v_i^d - b & \text{if } b \geq p_\epsilon \text{ and } b < b_j \text{ random bid } b_j, \\ v_i^d + v_i^p - b & \text{if } b \geq p_\epsilon \text{ and } b > b_j. \end{cases} \tag{6}$$

By introducing additional compensation, v_i^p, and, crucially, conditioning it on whether an agent's model is more compliant than that of another random agent, it becomes rational for agents to bid more than the price to obtain the minimum compliance threshold (unlike Theorem 1).

Incentivizing Agents to Build Compliant Models. We establish a guarantee that agents participating in Circa *maximize their utility with an optimal bid b_i^* that is larger than "the price required to attain ϵ compliance" (i.e., $b_i^* > p_\epsilon$)* in Theorem 2 below. Furthermore, agents bid in proportion to the value for additional compensation v_i^p that the regulator offers for extra-compliant models.

Theorem 2. *Agents participating in* Circa *Equation 6 follow an optimal bidding strategy $\hat{b}_i^*$ of,*

$$\hat{b}_i^* := p_\epsilon + v_i^p F(v_i^p) - \int_0^{v_i^p} F(z)dz \ > p_\epsilon, \quad u_i(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = v_i^d - p_\epsilon + \int_0^{v_i^p} F(z)dz, \tag{7}$$

where $F(\cdot)$ denotes the cumulative distribution function (CDF) of the random premium reward variable corresponding to the premium reward $v_i^p = V_i\lambda_i$ with $F(v_i^p) > 0$ for $v_i^p > 0$.

From Theorem 2, we find that each participating agent within Circa is expected to bid more than the price of compliance p_ϵ, thereby submitting a more compliant model than the required threshold.

Proposition 1 (Expected Participating Agent Bid)**.** *Let $\hat{b}_i^*$ denote the optimal bidding strategy from Theorem 2. The expected participating agent's bid over the distribution of premium rewards v_i^p, where $F(\cdot)$ and $f(\cdot)$ denote the CDF and probability density function (PDF) of v_i^p, is,*

$$\mathbb{E}[\hat{b}_i^*] = p_\epsilon + \int_0^{1/2} z f(z)(1 - F(z))dz \ > p_\epsilon. \tag{8}$$

Remark 2 (Outbidding Reserve Thresholding)**.** *In Theorem 2 and Proposition 1, participating agents submit models that are more compliant than the regulator's compliance threshold, $s_i^* = M^{-1}\big(b_i^*\big) > \epsilon$. This is an improvement over Theorem 1.*

Randomized and Deterministic Compliance Comparison. The standard version of CIRCA randomly pairs agents' models up against one another in a one-shot, zero-sum game. However, more deterministic pairing methods exist and are implementable within CIRCA. Such methods reduce the likelihood of unfair outcomes (*e.g.,* a scenario where the most compliant and second-most compliant models face off). The first alternative method is simple: repeated randomization. In this method, the randomization process is repeated r times and the number of "wins" for each agent w_i is stored by the regulator. Then, the regulator provides each agent i with a fraction of the premium reward: $(w_i/n) \cdot v_i^p$. In the case of an odd number of agents, each agent is held out h times (where $h = \lceil (r/n) \rceil$) while the remaining $n-1$ agents are paired up. The second alternative is deterministic. In this second method, the regulator computes the empirical CDF $\hat{F}(\cdot)$ across all n participating agents' model compliance bids and passes out $\hat{F}(b_i) \cdot v_i^p$ premium rewards to each agent i. For example, the participating agent with the median bid would receive half of the premium reward.

Remark 3 (Generalizable)**.** *Theorem 2 applies to any distribution for V_i and λ_i on $[0,1]$ and $[0,1/2]$, i.e., $V_i \sim \mathcal{D}_V(0,1)$ and $\lambda_i \sim \mathcal{D}_\lambda(0,1/2)$, respectively. Determining specific optimal bids, utility, and model compliance levels requires given distributions for V_i and λ_i.*

(Special Case 1) Uniform V_i and λ_i: Optimal Agent Strategy. Analysis of all-pay auctions (Amann & Leininger, 1996; Bhaskar, 2018; Tardos, 2017; Krishna, 2009), as well as other types of auctions, often assume a Uniform distribution for valuations. Therefore, our first analysis of CIRCA, below in Corollary 1, presumes Uniform distributions for V_i and λ_i. Corollary 1 determines that a participating agent's optimal strategy to maximize its utility is to submit a model with compliance levels larger than ϵ when their values V_i and λ_i come from a Uniform distribution.

Corollary 1 (Uniform Nash Bidding)**.** *Under Assumption 2, for agents having total value V_i and scaling factor λ_i both stemming from a Uniform distribution, with $v_i^d = (1-\lambda_i)V_i$, and $v_i^p = \lambda_i V_i$, their optimal bid and utility participating in* CIRCA *Equation 6 are $b_i^* := \min\{\hat{b}_i^*, 1\}$,*

$$\hat{b}_i^* = \begin{cases} p_\epsilon + \frac{(v_i^p)^2 \ln(p_\epsilon)}{p_\epsilon - 1} & \textit{if } 0 \leq v_i^p \leq \frac{p_\epsilon}{2}, \\ p_\epsilon + \frac{8(v_i^p)^2(\ln(2v_i^p) - 1/2) + p_\epsilon^2}{8(p_\epsilon - 1)} & \textit{if } \frac{p_\epsilon}{2} \leq v_i^p \leq \frac{1}{2}, \end{cases} \tag{9}$$

$$u_i(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = \begin{cases} \frac{2(v_i^p)^2 \ln(p_\epsilon)}{p_\epsilon - 1} + v_i^d - b_i^* & \textit{if } 0 \leq v_i^p \leq \frac{p_\epsilon}{2}, \\ \frac{2(v_i^p)^2(\ln(2v_i^p) - 1) + p_\epsilon}{p_\epsilon - 1} + v_i^d - b_i^* & \textit{if } \frac{p_\epsilon}{2} \leq v_i^p \leq \frac{1}{2}. \end{cases}$$

Participating agents submit models with compliance,

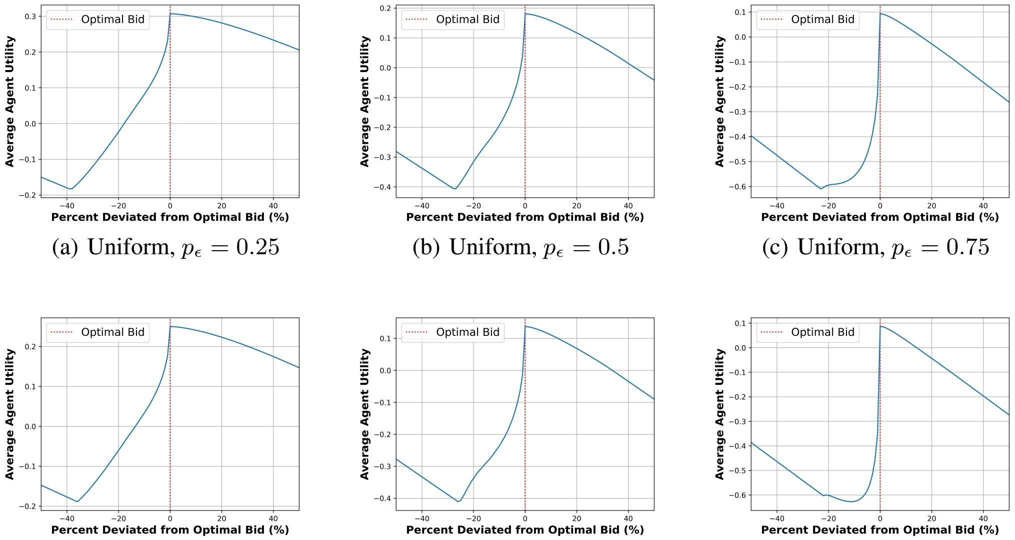
$$s_i^* := \begin{cases} M^{-1}(b_i^*) > \epsilon & \textit{if } u_i(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) > 0, \\ 0 \textit{ (no submission)} & \textit{else}. \end{cases} \tag{10}$$

(Special Case 2): Beta V_i and Uniform λ_i: Optimal Agent Strategy. In many instances, a realistic distribution for V_i is a Beta distribution with $\alpha, \beta = 2$. This distribution is Gaussian-like, with the bulk of the probability density placed in the middle. As such, it is realistic when agent values do not congregate amongst one another and outliers (near 0 or 1) are rare. The performance of CIRCA in this setting is analyzed in Corollary 2. Corollary 2 states that, under a Beta(2,2) distribution for V_i, agent i maximizes its utility with an optimal bid b_i^* larger than the price of ϵ compliance, $b_i^* > p_\epsilon$, resulting in a model above the ϵ-compliance threshold. Furthermore, Corollaries 1 and 2 surpass the baseline optimal bid $b_i^* = p_\epsilon$ for Reserve Thresholding (Theorem 1).

Corollary 2 (Beta Nash Bidding)**.** *Under Assumption 2, let agents have total value V_i and scaling factor λ_i stem from Beta $(\alpha, \beta = 2)$ and Uniform distributions respectively, with $v_i^d = (1-\lambda_i)V_i$ and $v_i^p = \lambda_i V_i$. Denote the CDF of the Beta distribution on $[0,1]$ as $F_\beta(x) = 3x^2 - 2x^3$. The optimal bid and utility for an agent participating in* CIRCA *(Equation 6) is $b_i^* := \min\{\hat{b}_i^*, 1\}$,*

$$\hat{b}_i^* = \begin{cases} p_\epsilon + \frac{3(v_i^p)^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} & 0 \leq v_i^p \leq \frac{p_\epsilon}{2}, \\ p_\epsilon + \frac{8(v_i^p)^2\left(6(v_i^p)^2 - 8v_i^p + 3\right) + p_\epsilon^3(3p_\epsilon - 4)}{8(1 - F_\beta(p_\epsilon))} & \frac{p_\epsilon}{2} \leq v_i^p \leq 1/2, \end{cases} \tag{11}$$

$$u(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = \begin{cases} v_i^d + \frac{6(v_i^p)^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} - b_i^* & 0 \leq v_i^p \leq \frac{p_\epsilon}{2}, \\ v_i^d + \frac{v_i^p\left(8(v_i^p)^3 - 12(v_i^p)^2 + 6v_i^p + p_\epsilon^2(2p_\epsilon - 3)\right)}{1 - F_\beta(p_\epsilon)} - b_i^* & \frac{p_\epsilon}{2} \leq v_i^p \leq 1/2. \end{cases} \tag{12}$$



(a) Uniform, $p_\epsilon = 0.25$ (b) Uniform, $p_\epsilon = 0.5$ (c) Uniform, $p_\epsilon = 0.75$

(d) Beta(2,2), $p_\epsilon = 0.25$ (e) Beta(2,2), $p_\epsilon = 0.5$ (f) Beta(2,2), $p_\epsilon = 0.75$

Figure 2: **Validation of Nash Bidding Equilibrium.** Agent utility is maximized when agents follow the theoretically optimal bidding function (red line) across different distributions and varying compliance prices $p_\epsilon \in \{0.25, 0.5, 0.75\}$. **Top row:** Uniform distribution results with the optimal bid from Equation 9. **Bottom row:** Beta($\alpha = \beta = 2$) distribution results with the optimal bid from Equation 11. *In all cases, agents attain strictly less utility when deviating from the optimal bid.*

Participating agents submit models with compliance,

$$s_i^* = \begin{cases} M^{-1}(b_i^*) > \epsilon & \textit{if } u_i(b_i^*;\, \boldsymbol{b}_{-i}, v_i^d, v_i^p) > 0, \\ 0 \textit{ (no submission)} & \textit{else}. \end{cases} \tag{13}$$

Remark 4 (Improved Utility & Participation)**.** *Through introduction of premium compensation, agent utility is improved (Equations 9 and 12) versus Reserve Thresholding (Equation 3). Thus, more agents achieve positive utility, thereby boosting overall agent utility and participation rate.*

The proofs of Theorems 1 and 2, Proposition 1, as well as Corollaries 1 and 2 are found within Appendix C. Since the premium compensation value v_i^p is a product of two random variables, the PDF and CDF of v_i^p becomes a piece-wise function (as shown within Appendix C). As a result, the optimal bidding and subsequent utility also becomes piece-wise in both Corollaries 1 and 2. Empirically, the correctness of the computed PDF and CDFs are verified within Appendix D.

6 Experiments

The results in Section 5 demonstrate that CIRCA creates incentives for any agents to submit compliant-exceeding models and to participate at rates higher than the baseline Reserve Thresholding mechanism in Section 4. Below, we validate these theoretical results empirically.

Lack of Baseline Regulatory Mechanisms. To the best of knowledge, there are no other comparable compliance mechanisms to regulate AI. As a result, the Reserve Threshold mechanism that is proposed in Section 4 is used as a baseline. While simple, the Reserve Threshold mechanism is a realistic baseline to compare against. For example, existing regulatory bodies, like the Environmental Protection Agency (EPA), follow similar steps before clearing products (*e.g.,* the EPA authorizes permits for discharging pollutants into water sources once water quality criteria are met).

Verifiable Nash Bidding Equilibria. In our first experiment, our goal is to validate that the theoretical bidding functions found in Corollaries 1 and 2 constitute Nash Equilibria. That is, an agent receives

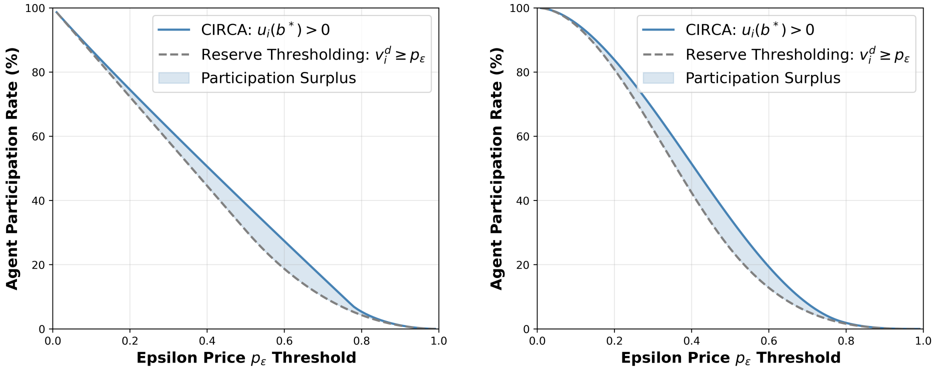


Figure 3: **Improved Participation with Uniform & Beta Values**. When total value stems from a (left) Uniform $V_i \sim U(0, 1)$ or (right) Beta distribution $V_i \sim \text{Beta}(\alpha = \beta = 2)$, agents participate at a higher rate in CIRCA than Reserve Thresholding.

worse utility if it deviates from this bidding strategy if other agents abide by it. To test this, we conduct a Monte Carlo simulation with $N = 100,000$ trials. In each trial, two agents, i and j, independently draw their valuations from either a Uniform or Beta(2,2) distribution (Figure 2). Each agent also receives a scaling factor λ_i that splits the total value into deployment $v_i^d = (1 - \lambda_i)V_i$ and premium compensation $v_i^p = \lambda_i V_i$ values. Once private values are provided, agents bid according to their optimal strategies for Uniform and Beta(2,2) distributions in Corollaries 1 and 2 respectively. Finally, we perturb agent i's optimal bid by $\pm 50\%$ and record the utility gained or lost by participating in CIRCA across the range of bids $b_i \in [b_i^*/2,\ 3b_i^*/2]$. We note that comparisons only occur if the other agent's bid is at least p_ϵ, in order to reflect how the auction mechanism in Algorithm 1 functions.

When inspecting Figure 2, it is clear that agent i's utility is maximized at its optimal bid b_i^*. Deviating from b_i^* results in less utility gained for agent i. Thus, *our Monte Carlo simulations validate our Nash Equilibria claims*: the Uniform and Beta optimal-bidding functions in Corollaries 1 and 2 indeed maximize agent utility in Figure 2.

Agent i's utility decays much quicker when bidding below its optimal bid b_i^*, since agents are **(i)** less likely to win the premium reward and **(ii)** at risk of losing the value from deployment if the bid does not reach p_ϵ. At a certain point, utility increases linearly once the agent continuously fails to bid p_ϵ. The linear improvement stems from the agent saving the cost of its bid, $-b_i$, shown in Equation 6.

Improved Agent Participation and Bid Size. In our second experiment, we showcase that agents participate at a higher rate and have a larger expected bid when participating in CIRCA compared to our Reserve Thresholding baseline, irrespective of the compliance threshold.

Agents participate when they can gain positive utility. For our Reserve Thresholding baseline this is simple, an agent i participates when their deployment value outstrips the price of compliance $v_i^d > p_\epsilon$. Thus, the participation rate for Reserve Thresholding is simply this probability $\mathbb{P}(v_i^d > p_\epsilon)$ across varying p_ϵ. In CIRCA, this probability, via Equation 7, is $\mathbb{P}(v_i^d + \int_0^{v_i^p} F(z)dz > p_\epsilon)$ across varying p_ϵ. We approximate these probabilities for both Uniform and Beta(2,2) distributions and plot the results in Figure 3. As expected, one can see that CIRCA achieves a greater participation rate, by upwards of 15%, compared to the Reserve Thresholding baseline across all values of p_ϵ. We note that participation rates are similar between the two methods for very low or high values of p_ϵ. The reason is simple. As $p_\epsilon \to 0$, all agents will achieve positive utility, and thus participate, irrespective of the mechanism. Likewise, as $p_\epsilon \to 1$, no agents will be able to achieve positive utility, and thus no agents will participate irrespective of the mechanism.

To analyze the expected participating bid in CIRCA, we leverage the results of Proposition 1. Namely, we approximate the integral in Equation 8, for both Uniform and Beta(2,2) distributions, to compute the expected agent bid for various values of p_ϵ. The results, provided in Figure 5, show that CIRCA incentivizes agents to bid more than they would than under our Reserve Thresholding baseline, by upwards of 20%, across all values of p_ϵ.

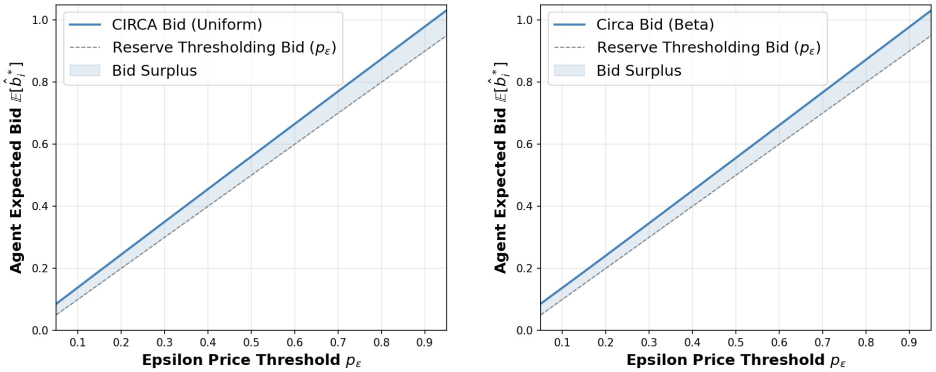


Figure 5: **Improved Compliance with Uniform & Beta Values**. When total value stems from a (left) Uniform $V_i \sim U(0,1)$ or (right) Beta distribution $V_i \sim \text{Beta}(\alpha = \beta = 2)$, agents bid more compliant models in CIRCA than Reserve Thresholding.

Compliance-Cost Case Study. Finally, a case study is conducted to demonstrate that in realistic settings, compliance is mapped to cost in a monotonically increasing way (as detailed in Assumption 2). While there are many compliance metrics to consider when gauging AI deployment, model fairness is analyzed, via equalized odds, for image classification in this study. Equalized odds measures if different groups have similar true positive rates and false positive rates (lower is better). Multiple VGG-16 models are trained on the Fairface dataset (Karkkainen & Joo, 2021) for fifty epochs (repeated ten times with different random seeds), and consider a gender classification task with race as the sensitive attribute. Models with the largest validation classification accuracy during training are selected for testing. Many types of costs exist for training compliant models, such as extensive architecture and hyper-parameter search. In this study, the cost of an agent acquiring more minority class data is considered. Acquiring more minority class data leads to a larger and more balanced dataset. Various mixtures of training data are simulated, starting from a 95:5 skew and scaling up to fully balanced training data with respect to the sensitive attribute. In this study, equalized odds performance is gauged on well-balanced test data for the models trained on various mixtures of data. The results of this case study are shown in Figure 4 and Table 3 (Appendix D).

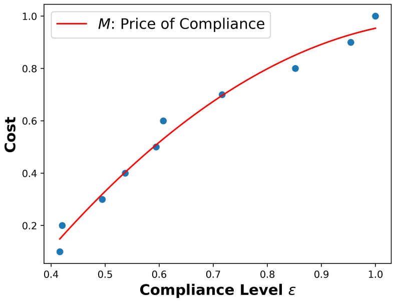


Figure 4: **Strictly Monotonic Compliance-Cost Relationship.** As the percentage of minority class data increases (greater cost), equalized odds metric improves (greater compliance) on Fairface.

As expected, in Table 3, the equalized odds score decreases (more compliant model) when collecting more minority class data (increased cost). To adjust equalized odds to fit into the setting where $\epsilon \in (0,1)$, the original equalized odds score are inverted and normalized. In Figure 4, one can see that compliance level is indeed monotonically increasing with respect to the cost.

7 Conclusion

As AI models grow, the risks associated with their misuse become significant, particularly given their opaque, black-box nature. Establishing robust algorithmic safeguards is crucial to protect users from unethical, unsafe, or illegally-deployed models. In this paper, we present a regulatory framework designed to ensure that only models deemed compliant by a regulator can be deployed for public use. Our key contribution is the development of an auction-based regulatory mechanism that simultaneously (i) enforces compliance standards and (ii) provably incentivizes agents to exceed minimum compliance thresholds. This approach encourages broader participation and the development of more compliant models compared to baseline regulatory methods. Empirical results confirm that our mechanism increases agent participation by 15% and raises agent spending on compliance by 20%, demonstrating its effectiveness to promote more compliant AI deployment.

ACKNOWLEDGMENTS

Bornstein, Che, Julapalli, and Huang are supported by DARPA Transfer from Imprecise and Abstract Models to Autonomous Technologies (TIAMAT) 80321, DARPA HR001124S0029-AIQ-FP-019, DOD-AFOSR-Air Force Office of Scientific Research under award number FA9550-23-1-0048, National Science Foundation TRAILS Institute (2229885). Private support was provided by Peraton and Open Philanthropy. The Authors acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot for contributing to this research result.

REFERENCES

EU Artificial Intelligence Act. The eu artificial intelligence act. *European Union*, 2024.

Erwin Amann and Wolfgang Leininger. Asymmetric all-pay auctions with incomplete information: the two-player case. *Games and economic behavior*, 14(1):1–18, 1996.

Dipto Barman, Ziyi Guo, and Owen Conlan. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 16:100545, 2024. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2024.100545. URL https://www.sciencedirect.com/science/article/pii/S2666827024000215.

Michael R Baye, Dan Kovenock, and Casper G De Vries. The all-pay auction with complete information. *Economic Theory*, 8:291–305, 1996.

V. Bhaskar. Lecture 8: All pay auction, January 2018.

Jack Brewster. How i built an ai-powered, self-running propaganda machine for $105. *The Wall Street Journal*, 2024. URL https://www.wsj.com/politics/how-i-built-an-ai-powered-self-running-propaganda-machine-for-105-e9888705.

Alexandra Bruell. New york times sues microsoft and openai, alleging copyright infringement. *The Wall Street Journal*, 2023. URL https://www.wsj.com/tech/ai/new-york-times-sues-microsoft-and-openai-alleging-copyright-infringement-fd85e1c4.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

C. Coglianese and R.A. Kagan. *Regulation and Regulatory Processes*. International library of essays in law and society. Ashgate, 2007. ISBN 9780754625186. URL https://books.google.com/books?id=C0RnNQAACAAJ.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

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.

Dominic DiPalantino and Milan Vojnovic. Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pp. 119–128, 2009.

Ian Gemp, Thomas Anthony, Janos Kramar, Tom Eccles, Andrea Tacchetti, and Yoram Bachrach. Designing all-pay auctions using deep learning and multi-agent simulation. *Scientific Reports*, 12 (1):16937, 2022.

Jacob K Goeree and John L Turner. All-pay-all auctions. *University of Virginia, Mimeo*, 2000.

Lawrence H Goulder and Andrew R Schein. Carbon taxes versus cap and trade: a critical review. *Climate Change Economics*, 4(03):1350010, 2013.

The White House. Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence, 2023.

Nikolaus Howe, Ian McKenzie, Oskar Hollinsworth, Michał Zajac, Tom Tseng, Aaron Tucker, Pierre-Luc Bacon, and Adam Gleave. Effects of scale on language model robustness. *arXiv preprint arXiv:2407.18213*, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

California Legislative Information. Sb-1047 safe and secure innovation for frontier artificial intelligence models act. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047, 2024.

Denny Jacob. Openai forms new committee to evaluate safety, security. *The Wall Street Journal*, 2024. URL https://www.wsj.com/tech/ai/openai-forms-new-committee-to-evaluate-safety-security-4a6e74bb.

Meena Jagadeesan, Michael I Jordan, and Jacob Steinhardt. Safety vs. performance: How multi-objective learning reduces barriers to market entry. *arXiv preprint arXiv:2409.03734*, 2024.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Vijay Krishna. *Auction theory*. Academic press, 2009.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*, 2024.

Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson, and Nico Grant. How tech giants cut corners to harvest data for a.i. *The New York Times*, 2024. URL https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

J. Edward Moreno. Boom in a.i. prompts a test of copyright law. *The New York Times*, 2023. URL https://www.nytimes.com/2023/12/30/business/media/copyright-law-ai-media.html.

Gary D Lopez Munoz, Amanda J Minnich, Roman Lutz, Richard Lundeen, Raja Sekhar Rao Dheekonda, Nina Chikanov, Bolor-Erdene Jagdagdorj, Martin Pouliot, Shiven Chawla, Whitney Maxwell, et al. Pyrit: A framework for security risk identification and red teaming in generative ai system. *arXiv preprint arXiv:2410.02828*, 2024.

Terrence Neumann, Sooyong Lee, Maria De-Arteaga, Sina Fazelpour, and Matthew Lease. Diverse, but divisive: Llms can exaggerate gender differences in opinion related to harms of misinformation. *arXiv preprint arXiv:2401.16558*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Mark R Powell. *Science at EPA: Information in the regulatory process*. Routledge, 2014.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Adi Robertson. Google apologizes for ‘missing the mark’ after gemini generated racially diverse nazis. *The Verge*, 2024. URL https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical.

Manuel J García Rodríguez, Vicente Rodríguez-Montequín, Pablo Ballesteros-Pérez, Peter ED Love, and Regis Signor. Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, 133:104047, 2022.

Deepa Seetharaman. Openai, meta and google sign on to new child exploitation safety measures. *The Wall Street Journal*, 2024. URL https://www.wsj.com/tech/ai/ai-developers-agree-to-new-safety-measures-to-fight-child-exploitation-2a58129c.

Ron Siegel. All-pay contests. *Econometrica*, 77(1):71–92, 2009.

Robert N Stavins. A meaningful us cap-and-trade system to address climate change. *Harv. Envtl. L. Rev.*, 32:293, 2008.

Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*, 2024.

Eva Tardos. Lecture 17: All-pay auctions, March 2017. URL https://www.cs.cornell.edu/courses/cs6840/2017sp/lecnotes/lec17.pdf.

U.S. Food and Drug Administration. E19 a selective approach to safety data collection in specific late-stage pre-approval or post-approval clinical trials, 2022. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e19-selective-approach-safety-data-collection-specific-late-stage-pre-approval-or-post-approval.

Gail A Van Norman. Drugs, devices, and the fda: part 1: an overview of approval processes for drugs. *JACC: Basic to Translational Science*, 1(3):170–179, 2016.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

Jeremy White. How strangers got my email address from chatgpt’s model. *The New York Times*, 2023. URL https://www.nytimes.com/interactive/2023/12/22/technology/openai-chatgpt-privacy-exploit.html.

Jeremy White. See how easily a.i. chatbots can be taught to spew disinformation. *The New York Times*, 2024. URL https://www.nytimes.com/interactive/2024/05/19/technology/biased-ai-chatbots.html.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Mohammad Yaghini, Patty Liu, Franziska Boenisch, and Nicolas Papernot. Regulation games for trustworthy machine learning. *arXiv preprint arXiv:2402.03540*, 2024.

Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, and Amelia Glaese. Trading inference-time compute for adversarial robustness. https://cdn.openai.com/papers/trading-inference-time-compute-for-adversarial-robustness-20250121_1.pdf, 2025.

Table 1: Comparison of AI Regulation Frameworks

Feature	**CIRCA (This Paper)**	Jagadeesan et al. (2024)	Yaghini et al. (2024)	**All-Pay Auctions (General)**
Overview	Formulates AI regulation as an auction to derive Nash Equilibria.	Penalizes larger companies more for unsafe AI models.	Introduces a multi-agent, multi-objective regulatory game.	Diverse formulations of all-pay auctions.
Regulatory Scheme?	✓	✗	✓	✗
Compliance-Aware?	✓	✗	✓	✗
Theoretical Guarantees?	✓	✗	✗	✓
Incentivizing Over-Compliance?	✓	✗	✗	✗
Multiple Model Builders?	✓	✓	✗	✓
Single Round (Simple)?	✓	✓	✗	✓

A Notation Table

Table 2: Notating and Defining all Variables Listed Within CIRCA.

Definition	**Notation**
Regulator	R
Number of Agents	n
Compliance Threshold	ϵ
Compliance-to-Cost Function	M
Price of Attaining Compliance	p_ϵ
Agent i Bid	b_i
Agent i's Optimal Bid	b_i^*
All Other Agents Bids	$\boldsymbol{b}_{-i}$
Agent i Utility	u_i
Agent i Model Compliance	s_i
Total Value for Agent i	V_i
Total Value Distribution	$\mathcal{D}_V$
Agent i Scaling Factor	λ_i
Scaling Factor Distribution	$\mathcal{D}_\lambda$
Deployment Value for Agent i	v_i^d
Premium Compensation Value for Agent i	v_i^p
Probability Density Function for Premium Compensation	f_v
Cumulative Distribution Function for Premium Compensation	F_v

B BINARY AND DISCRETE COMPLIANCE IN CIRCA

Our framework still works within binary or discrete settings. This is important when dealing with properties or metrics that are not continuous, like how the EU AI Act evaluates AI risk into minimal, limited, high, and unacceptable tiers (Act, 2024). The rationale behind why CIRCA works for binary or discrete settings is that models can still be ranked or compared against each other depending on how well they satisfy the given metric or property.

For example, models can be separated into Pass/Fail categories, where the Pass category can be further split into High/Medium/Low sub-categories. All models achieving at least Low Pass compliance are cleared for deployment. While a model either complies or does not, the models can still be gauged on how well they comply (e.g., High/Medium/Low). Since a ranking of models can still be generated, premium rewards can be provided to higher-passing models.

In situations where the regulatory policy is black and white, for example "your model must be trained with differential privacy", CIRCA still holds as an ordering or ranking between models can still be ascertained. In the example of differential privacy, *any* model that is trained with differential privacy would be cleared for deployment. However, it is also true that differential privacy can be gauged by a level of privacy ϵ_{DP} (not to be confused with our compliance threshold ϵ). Models with smaller values of ϵ_{DP} will be provided additional premium rewards since they are more compliant (*i.e.,* more private). Thus, CIRCA would still incentivize agents to become more private even when there is a binary compliance metric.

C THEORETICAL PROOFS

Below, we provide the full proofs of our Theorems and Corollaries presented within our work.

C.1 PROOF OF THEOREM 1

Theorem 1 (Restated)**.** *Under Assumption 2, agents participating in Reserve Thresholding Equation 2 have an optimal bid and utility of,*

$$b_i^* = p_\epsilon, \quad u_i(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d) = v_i^d - p_\epsilon,$$

and submit models with the following compliance level,

$$s_i^* = \begin{cases} \epsilon & \textit{if } u_i(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d) > 0, \\ 0 \textit{ (no submission)} & \textit{else.} \end{cases}$$

Proof. From agent i's utility within Reserve Thresholding, Equation 2, it is clear that $u_i(0) = 0$. We proceed to break the proof up into cases where agents have (1) a deployment value equal to or less than the price of compliance p_ϵ and (2) a deployment value larger than p_ϵ.

Case 1: $v_i^d \leq p_\epsilon$. Leveraging Equation 2, when $b < p_\epsilon$ the indicator function returns $1_{(b \geq p_\epsilon)} = 0$. Therefore, agent utility is always negative when bidding less than p_ϵ,

$$u_i(b) = -b < 0. \tag{14}$$

When $b \geq p_\epsilon$, the indicator function returns $1_{(b \geq p_\epsilon)} = 1$. Thus, agent utility becomes,

$$u_i(b) = v_i^d - b < 0. \tag{15}$$

Since this function is strictly decreasing with respect to b, the function is maximized at the smallest bid $b = p_\epsilon$,

$$b_i^* = \underset{b \in [p_\epsilon, 1]}{\arg\max}\, u_i(b) = p_\epsilon \longrightarrow u_i(p_\epsilon) = v_i^d - p_\epsilon \leq p_\epsilon - p_\epsilon = 0. \tag{16}$$

For an agent with deployment value at most equal to p_ϵ, the upper bound on attainable utility when it participates, *i.e.,* $b \in (0, 1]$, is zero (Equation 16). Thus, agents have nothing to gain by participating, as they already start at zero utility $u_i(0) = 0$. As a result, agents will not submit a model, $s_i^* = M(0) = 0$.

Case 2: $v_i^d > p_\epsilon$. By the same two-region argument as Case 1, bidding $b < p_\epsilon$ results in negative utility. Similarly, bidding $b \geq p_\epsilon$ results in an agent optimally bidding $b_i^* = p_\epsilon$. However, in Case 2, the agent utility becomes,

$$u_i(b_i^*) = v_i^d - p_\epsilon > 0. \tag{17}$$

An agent with deployment value larger than p_ϵ will have a non-negative optimal utility when it participates (Equation 17). Furthermore, at this optimal bid, the corresponding compliance level is $s_i^* = M^{-1}(p_\epsilon) = \epsilon$. □

C.2 Proof of Theorem 2

Theorem 2 (Restated). *Agents participating in* Circa *Equation 6 will follow an optimal bidding strategy $\hat{b}_i^*$ of,*

$$\hat{b}_i^* := p_\epsilon + v_i^p F(v_i^p) - \int_0^{v_i^p} F(z)dz > p_\epsilon,$$

where $F(\cdot)$ denotes the cumulative distribution function of the random premium reward variable corresponding to the premium reward $v_i^p = V_i\lambda_i$ with $F(v_i^p) > 0$ for $v_i^p > 0$.

Proof. Before beginning our proof, we note that each agent i cannot alter its own valuation v_i^p for winning the all-pay auction. Each valuation is private (unknown by other agents) and predetermined: total reward V_i and partition factor λ_i are randomly selected from a given distribution $\mathcal{D}$ on $[0, 1]$ and $[0, 1/2]$ respectively by "nature". We define the cumulative distribution function for the auction reward $v_i^p = V_i\lambda_i$ as $F(\cdot)$ and the probability distribution function as $f(\cdot)$. From Equation 6, we find that not participating (*i.e.,* $b = 0$) results in no utility,

$$u_i(0) = 0. \tag{18}$$

An agent receives negative utility if its bid does not reach the price of compliance p_ϵ,

$$\max_{b\in(0,p_\epsilon)} u_i(b) < 0. \tag{19}$$

Consequently, rational agents will either opt not to participate (notated as the set of agents N) or participate (notated as the set of agents P) and bid at least p_ϵ. We define these groups as,

$$N = \{i \in [n] \mid \max_{b\in[0,1]} u_i(b) \leq 0\}, \tag{20}$$

$$P = \{i \in [n] \mid \max_{b\in[0,1]} u_i(b) > 0\}. \tag{21}$$

From here, we only focus on agents $i \in P$ which participate (*i.e.,* have utility to be gained by participating). As a result from Equations 18 and 19, Equation 21 transforms into,

$$P = \{i \in [n] \mid \max_{b\in[p_\epsilon,1]} u_i(b_i) > 0\}. \tag{22}$$

The result of Equation 22 is that participating agents bid at least p_ϵ. This is important, as every participating agent knows that all rival agents j they will possibly be compared against have $b_j \in [p_\epsilon, 1]$. Agents can dictate how much they bid, and we design our auction to ensure that agents bid in proportion to their valuation.

Following previous literature (Amann & Leininger, 1996; Bhaskar, 2018; Tardos, 2017; Krishna, 2009), we desire a *monotone increasing equilibrium* bidding function $\beta(v^p) : [0, 1/2] \to [p_\epsilon, 1]$. We conjecture that all rival agents $j \neq i$ follow this bidding function $\beta(v_j^p)$. We then derive agent i's best response under this conjecture, and verify that the best response is $\beta(v_i^p)$, establishing Nash equilibrium by symmetry. Using a bidding function transforms agent utility, for a given bid b, into the following,

$$\begin{aligned} u_i(b;\ v_i^d, v_i^p) &= \big(v_i^d + v_i^p \cdot 1_{(\text{if } i \text{ wins auction})}\big) \cdot \underbrace{1_{(\text{if } b \geq p_\epsilon)}}_{\text{satisfied for agents } i\in P} - \beta(b), \\ &= \mathbb{P}\big(\beta(b) > \beta(v_j^p)\big)v_i^p - \beta(b) + v_i^d, \quad b_j \sim \text{randomly sampled agent bid.} \end{aligned} \tag{23}$$

Since $\beta(\cdot)$ is monotone increasing up to 1, and agents bidding $b = 1$ automatically win, one can see that $\mathbb{P}\big(\beta(b) > \beta(v_j^p)\big) = \mathbb{P}\big(b > v_j^p\big)$. Since $v_j^p \sim F$, it follows that $\mathbb{P}\big(b > v_j^p\big) = F(b)$. Thus, the utility function above becomes,

$$u_i(b;\ v_i^d, v_i^p) = v_i^p F(b) - \beta(b) + v_i^d. \tag{24}$$

At equilibrium, truthful reporting $b = v_i^p$ must be optimal. Differentiating with respect to b and solving the first order conditions at $b = v_i^p$ yields,

$$\frac{\partial u_i}{\partial b}\bigg|_{b=v_i^p} = v_i^p f(v_i^p) - \beta'(v_i^p) = 0. \tag{25}$$

This yields the following ODE (as v_i^p takes on values between 0 and 1/2),

$$\beta'(v) = v f(v), \quad \text{for } v \in [0, 1/2]. \tag{26}$$

Now, solving the ODE at the specific value v_i^p,

$$\int_0^{v_i^p} \beta'(v)dv = \int_0^{v_i^p} v f(v) dv. \tag{27}$$

From Equation 22, all participating agents bid at a minimum $\beta(0) = p_\epsilon$. Therefore, $\beta(0) = p_\epsilon$ is our boundary condition. Using integration by parts, and applying the boundary condition yields our desired result,

$$\beta(v_i^p) - \beta(0) = vF(v)\big|_0^{v_i^p} - \int_0^{v_i^p} F(z)dz, \tag{28}$$

$$\hat{b}_i^* = \beta(v_i^p) := p_\epsilon + v_i^p F(v_i^p) - \int_0^{v_i^p} F(z)dz. \tag{29}$$

Since $\int_0^{v_i^p} F(z)dz < \int_0^{v_i^p} F(v_i^p)dz = v_i^p F(v_i^p)$ for $v_i^p > 0$, it follows that,

$$\hat{b}_i^* = p_\epsilon + v_i^p F(v_i^p) - \int_0^{v_i^p} F(z)dz > p_\epsilon. \tag{30}$$

The general utility for CIRCA is detailed in Equation 5 as $u_i(b;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = v_i^d - p_\epsilon + v_i^p \cdot \mathbf{1}[\text{agent } i \text{ wins}]$. This is represented more formally as,

$$u_i(b;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = v_i^d - b + v_i^p F(v_i^p). \tag{31}$$

Substituting in the optimal bid yields,

$$u_i(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = v_i^d - \big(p_\epsilon + v_i^p F(v_i^p) - \int_0^{v_i^p} F(z)dz\big) + v_i^p F(v_i^p), \tag{32}$$

$$= v_i^d - p_\epsilon + \int_0^{v_i^p} F(z)dz. \tag{33}$$

□

C.3 Proof of Proposition 1

Proposition 1 (Restated)**.** *Let $\hat{b}_i^*$ denote the optimal bidding strategy from Theorem 2. The expected participating agent's bid over the distribution of premium rewards v_i^p, where $F(\cdot)$ and $f(\cdot)$ denote the CDF and probability density function (PDF) of v_i^p, is,*

$$\mathbb{E}[\hat{b}_i^*] = p_\epsilon + \int_0^{1/2} z f(z)(1 - F(z))dz \ > p_\epsilon. \tag{34}$$

Proof. From Theorem 2, the optimal bid is,

$$\hat{b}_i^* = p_\epsilon + v_i^p F(v_i^p) - \int_0^{v_i^p} F(z)\, dz. \tag{35}$$

We simplify $v_i^p F(v_i^p) - \int_0^{v_i^p} F(z)\, dz$ using integration by parts. Recall that for any function F,

$$\int_0^v z f(z)\, dz = \Big[zF(z)\Big]_0^v - \int_0^v F(z)\, dz = vF(v) - \int_0^v F(z)\, dz. \tag{36}$$

Therefore,

$$\hat{b}_i^* = p_\epsilon + \int_0^{v_i^p} z f(z)\, dz. \tag{37}$$

Taking expectations over v_i^p, and applying the Law of the Unconscious Statistician, yields

$$\mathbb{E}[\hat{b}_i^*] = p_\epsilon + \mathbb{E}\left[\int_0^{v_i^p} z f(z)\, dz\right] = p_\epsilon + \int_0^{1/2} \left(\int_0^v z f(z)\, dz\right) f(v)\, dv, \tag{38}$$

where the outer $f(v)$ is the PDF of v_i^p and the domain upper limit is $1/2$ since $v_i^p = V_i \lambda_i \leq 1/2$ by construction. The double integral is taken over the region $\{(z, v) : 0 \leq z \leq v \leq 1/2\}$, which is a triangle in the (z, v) plane with vertices $(0, 0)$, $(0, 1/2)$, and $(1/2, 1/2)$. Since $zf(z)f(v) \geq 0$ on this region, we can switch the order of integration (via Fubini's theorem). For a fixed z, v ranges over $[z, 1/2]$, giving,

$$\mathbb{E}[\hat{b}_i^*] = p_\epsilon + \int_0^{1/2} z f(z) \left(\int_z^{1/2} f(v)\, dv\right) dz. \tag{39}$$

The inner integral evaluates (since $F(1/2) = 1$ as $v_i^p \leq 1/2$) to,

$$\int_z^{1/2} f(v)\, dv = F(1/2) - F(z) = 1 - F(z), \tag{40}$$

This leads to our final result,

$$\mathbb{E}[\hat{b}_i^*] = p_\epsilon + \int_0^{1/2} z f(z)(1 - F(z))\, dz. \tag{41}$$

□

C.4 Proof of Corollary 1

Corollary 1 (Restated). *Under Assumption 2, for agents having total value V_i and scaling factor λ_i both stemming from a Uniform distribution, with $v_i^d = (1 - \lambda_i)V_i$, and $v_i^p = \lambda_i V_i$, their optimal bid and utility participating in* CIRCA *Equations 6 are $b_i^* := \min\{\hat{b}_i^*, 1\}$,*

$$\hat{b}_i^* = \begin{cases} p_\epsilon + \frac{(v_i^p)^2 \ln(p_\epsilon)}{p_\epsilon - 1} & \textit{if } 0 \leq v_i^p \leq \frac{p_\epsilon}{2}, \\ p_\epsilon + \frac{8(v_i^p)^2(\ln(2v_i^p) - 1/2) + p_\epsilon^2}{8(p_\epsilon - 1)} & \textit{if } \frac{p_\epsilon}{2} \leq v_i^p \leq \frac{1}{2}, \end{cases}$$

$$u_i(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = \begin{cases} \frac{2(v_i^p)^2 \ln(p_\epsilon)}{p_\epsilon - 1} + v_i^d - b_i^* & \textit{if } 0 \leq v_i^p \leq \frac{p_\epsilon}{2}, \\ \frac{2(v_i^p)^2(\ln(2v_i^p) - 1) + p_\epsilon}{p_\epsilon - 1} + v_i^d - b_i^* & \textit{if } \frac{p_\epsilon}{2} \leq v_i^p \leq \frac{1}{2}. \end{cases}$$

Participating agents submit models with compliance,

$$s_i^* := \begin{cases} M^{-1}(b_i^*) > \epsilon & \textit{if } u_i(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) > 0, \\ 0 \textit{ (no submission)} & \textit{else}. \end{cases}$$

Proof. Let $v_i^p := V_i \lambda_i$, where $V_i \sim U[p_\epsilon, 1]$ and $\lambda_i \sim U[0, 1/2]$. The reason that V_i is within the interval $[p_\epsilon, 1]$, is that all participating agents must have a value of at least p_ϵ or else they would not have rationale to bid. The smallest value of V_i such that this is possible is p_ϵ, so it is the lower bound on this interval. Our first goal is to find the PDF of v_i^p, $f_{v_i^p}(\cdot)$.

We begin solving for $f_{v_i^p}(\cdot)$ by using a change of variables. For the product of two random variables $v = x_1 \cdot x_2$, let $y_1 = x_1 \cdot x_2$ and $y_2 = x_2$. Thus, we find inversely that $x_2 = y_2$ and $x_1 = y_1/y_2$. Since x_1 and x_2 are independent and both uniform, we find that,

$$f_{y_1,y_2}(x_1, x_2) = (\frac{1}{1-p_\epsilon})(\frac{1}{1/2-0}) = \frac{2}{1-p_\epsilon}. \tag{42}$$

When using the change of variables this becomes,

$$f_{y_1,y_2}(y_1, y_2) = f_{y_1,y_2}(x_1, x_2)|J| = \frac{2}{(1-p_\epsilon)y_2}, \quad |J| = \left| \begin{pmatrix} 1/y_2 & -y_1/y_2^2 \\ 0 & 1 \end{pmatrix} \right| = 1/y_2 \tag{43}$$

Marginalizing out y_2 (a non-negative value) yields,

$$f_{y_1}(y_1) = \int_0^\infty \frac{2}{(1-p_\epsilon)y_2} dy_2. \tag{44}$$

The bounds of integration depend upon the value of y_1. The change of variable to the (y_1, y_2) space, where $0 \le y_1, y_2 \le 1/2$, results in a new region of possible variable values. This region is a triangle bounded by the three vertices: $(0, 0)$, $(p_\epsilon/2, 1/2)$, and $(1/2, 1/2)$. Thus, the bounds of marginalization depend upon the value of y_1. For $0 \le y_1 \le p_\epsilon/2$ we have,

$$f_{y_1}(y_1) = \int_{y_1}^{y_1/p_\epsilon} \frac{2}{(1-p_\epsilon)y_2} dy_2 = \frac{2}{(1-p_\epsilon)} [\ln(y_2)|_{y_1}^{y_1/p_\epsilon}] = \frac{2\ln(p_\epsilon)}{(p_\epsilon - 1)}. \tag{45}$$

For $p_\epsilon/2 \le y_1 \le 1/2$ we have,

$$f_{y_1}(y_1) = \int_{y_1}^{1/2} \frac{2}{(1-p_\epsilon)y_2} dy_2 = \frac{2}{(1-p_\epsilon)} [\ln(y_2)|_{y_1}^{1/2}] = \frac{2\ln(2y_1)}{(p_\epsilon - 1)}. \tag{46}$$

Thus, as a piecewise function the PDF is formally,

$$f_{y_1}(y_1) = \begin{cases} \frac{2\ln(p_\epsilon)}{(p_\epsilon - 1)} & \text{for } 0 \le y_1 \le \frac{p_\epsilon}{2}, \\ \frac{2\ln(2y_1)}{(p_\epsilon - 1)} & \text{for } \frac{p_\epsilon}{2} \le y_1 \le 1/2. \end{cases} \tag{47}$$

Now, the CDF is determined through integration,

$$F_{y_1}(y_1) = \int_0^{y_1} f_{y_1}(y_1) dy_1 = \begin{cases} \frac{2y_1\ln(p_\epsilon)}{(p_\epsilon - 1)} & \text{for } 0 \le y_1 \le \frac{p_\epsilon}{2}, \\ \frac{2y_1(\ln(2y_1)-1)+p_\epsilon}{(p_\epsilon - 1)} & \text{for } \frac{p_\epsilon}{2} \le y_1 \le 1/2. \end{cases} \tag{48}$$

We can integrate the CDF to get,

$$\int_0^{y_1} F_{y_1}(y_1) = \begin{cases} \frac{y_1^2\ln(p_\epsilon)}{(p_\epsilon - 1)} & \text{for } 0 \le y_1 \le \frac{p_\epsilon}{2}, \\ \frac{4y_1^2(2\ln(2y_1)-3)+8y_1p_\epsilon - p_\epsilon^2}{8(p_\epsilon - 1)} & \text{for } \frac{p_\epsilon}{2} \le y_1 \le 1/2. \end{cases} \tag{49}$$

Plugging all of this back into Equation 7 yields,

$$\begin{aligned} \hat{b}_i^* &= \begin{cases} p_\epsilon + v_i^p \frac{2v_i^p\ln(p_\epsilon)}{p_\epsilon - 1} - \frac{(v_i^p)^2\ln(p_\epsilon)}{p_\epsilon - 1}, \\ p_\epsilon + v_i^p \frac{2v_i^p(\ln(2v_i^p)-1)+p_\epsilon}{(p_\epsilon - 1)} - \frac{4(v_i^p)^2(2\ln(2v_i^p)-3)+8v_i^pp_\epsilon - p_\epsilon^2}{8(p_\epsilon - 1)}, \end{cases} \\ &= \begin{cases} p_\epsilon + \frac{(v_i^p)^2\ln(p_\epsilon)}{p_\epsilon - 1} & \text{if } 0 \le v_i^p \le \frac{p_\epsilon}{2}, \\ p_\epsilon + \frac{8(v_i^p)^2(\ln(2v_i^p)-1/2)+p_\epsilon^2}{8(p_\epsilon - 1)} & \text{if } \frac{p_\epsilon}{2} \le v_i^p \le \frac{1}{2}. \end{cases} \end{aligned} \tag{50}$$

Since b_i cannot be larger than 1, we cap the bidding function at one via,

$$b_i^* := \min\{\hat{b}_i^*, 1\}. \tag{51}$$

The utility gained by agent i for using such a bidding function is,

$$u(b_i^*) = \begin{cases} v_i^d - b_i^* + \frac{2(v_i^p)^2\ln(p_\epsilon)}{p_\epsilon - 1} & \text{for } 0 \le v_i^p \le \frac{p_\epsilon}{2}, \\ v_i^d - b_i^* + \frac{2(v_i^p)^2(\ln(2v_i^p)-1)+p_\epsilon}{(p_\epsilon - 1)} & \text{for } \frac{p_\epsilon}{2} \le v_i^p \le 1/2. \end{cases} \tag{52}$$

When this utility is larger than 0, the agent will participate otherwise the agent will not submit a model to the regulator. Finally, we can find the optimal compliance level by using Assumption 2,

$$s_i^* := M^{-1}(b_i^*). \tag{53}$$

□

C.5 Proof of Corollary 2

Corollary 2 (Restated)**.** *Under Assumption 2, let agents have total value V_i and scaling factor λ_i stem from Beta ($\alpha = \beta = 2$) and Uniform distributions respectively, with $v_i^d = (1 - \lambda_i)V_i$ and $v_i^p = \lambda_i V_i$. Denote the CDF of the Beta distribution on $[0, 1]$ as $F_\beta(x) = 3x^2 - 2x^3$. The optimal bid and utility for agents participating in* Circa *Equation 6 are,*

$$b_i^* := \min\{\hat{b}_i^*, 1\}, \quad \hat{b}_i^* = \begin{cases} p_\epsilon + \frac{3(v_i^p)^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} & \textit{if } 0 \le v_i^p \le \frac{p_\epsilon}{2}, \\ p_\epsilon + \frac{8(v_i^p)^2\left(6(v_i^p)^2 - 8v_i^p + 3\right) + p_\epsilon^3(3p_\epsilon - 4)}{8(1 - F_\beta(p_\epsilon))} & \textit{if } \frac{p_\epsilon}{2} \le v_i^p \le \frac{1}{2}, \end{cases}$$

$$u(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) = \begin{cases} v_i^d + \frac{6(v_i^p)^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} - b_i^* & \textit{for } 0 \le v_i^p \le \frac{p_\epsilon}{2}, \\ v_i^d + \frac{v_i^p\left(8(v_i^p)^3 - 12(v_i^p)^2 + 6v_i^p + p_\epsilon^2(2p_\epsilon - 3)\right)}{1 - F_\beta(p_\epsilon)} - b_i^* & \textit{for } \frac{p_\epsilon}{2} \le v_i^p \le 1/2. \end{cases}$$

Participating agents submit models with compliance,

$$s_i^* = \begin{cases} M^{-1}(b_i^*) > \epsilon & \textit{if } u_i(b_i^*;\ \boldsymbol{b}_{-i}, v_i^d, v_i^p) > 0, \\ 0 \textit{ (no model submission)} & \textit{else}. \end{cases}$$

Proof. Similar to Corollary 1, we begin solving for $f_{v_i^p}(\cdot)$ using a change of variables. For the product of two random variables $v = x_1 \cdot x_2$, let $y_1 = x_1 \cdot x_2$ and $y_2 = x_2$. Inversely, $x_2 = y_2$ and $x_1 = y_1/y_2$. While x_1 and x_2 are independent, x_1 comes from a Beta distribution and x_2 from a Uniform one. The PDF and CDF of a Beta distribution, with $\alpha = \beta = 2$, on $[0, 1]$ are defined as,

$$f_\beta(x) := 6x(1 - x), \tag{54}$$

$$F_\beta(x) := 3x^2 - 2x^3. \tag{55}$$

Now, the PDF over y_1, y_2 is defined as,

$$f_{y_1,y_2}(x_1, x_2) = \left(\frac{6x_1(1 - x_1)}{1 - F_\beta(p_\epsilon)}\right)\left(\frac{1}{1/2 - 0}\right) = \frac{12x_1(1 - x_1)}{1 - F_\beta(p_\epsilon)}. \tag{56}$$

When using the change of variables this becomes,

$$f_{y_1,y_2}(y_1, y_2) = f_{y_1,y_2}(x_1, x_2)|J| = \frac{12y_1(1 - \frac{y_1}{y_2})}{(1 - F_\beta(p_\epsilon))y_2^2}, \quad |J| = \left| \begin{pmatrix} 1/y_2 & -y_1/y_2^2 \\ 0 & 1 \end{pmatrix} \right| = 1/y_2 \tag{57}$$

Marginalizing out y_2 (a non-negative value) yields,

$$f_{y_1}(y_1) = \frac{12y_1}{1 - F_\beta(p_\epsilon)} \int_0^\infty \frac{1}{y_2^2} - \frac{y_1}{y_2^3} dy_2. \tag{58}$$

The bounds of integration depend upon the value of y_1. The change of variable to the (y_1, y_2) space, where $0 \le y_1, y_2 \le 1/2$, results in a new region of possible variable values. This region is a triangle bounded by the three vertices: $(0, 0)$, $(p_\epsilon/2, 1/2)$, and $(1/2, 1/2)$. Thus, the bounds of marginalization depend upon the value of y_1. For $0 \le y_1 \le p_\epsilon/2$ we have,

$$\begin{aligned} f_{y_1}(y_1) &= \frac{12y_1}{1 - F_\beta(p_\epsilon)} \int_{y_1}^{y_1/p_\epsilon} \frac{1}{y_2^2} - \frac{y_1}{y_2^3} dy_2 = \frac{12y_1}{1 - F_\beta(p_\epsilon)}[-\frac{1}{y_2} + \frac{y_1}{2y_2^2}\Big|_{y_1}^{y_1/p_\epsilon}] \\ &= \frac{12y_1}{1 - F_\beta(p_\epsilon)}[-\frac{p_\epsilon}{y_1} + \frac{p_\epsilon^2}{2y_1} + \frac{1}{y_1} - \frac{1}{2y_1}] = \frac{6(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)}. \end{aligned} \tag{59}$$

For $p_\epsilon/2 \le y_1 \le 1/2$ we have,

$$\begin{aligned} f_{y_1}(y_1) &= \frac{12y_1}{1 - F_\beta(p_\epsilon)} \int_{y_1}^{1/2} \frac{1}{y_2^2} - \frac{y_1}{y_2^3} dy_2 = \frac{12y_1}{1 - F_\beta(p_\epsilon)}[-\frac{1}{y_2} + \frac{y_1}{2y_2^2}\Big|_{y_1}^{1/2}] \\ &= \frac{12y_1}{1 - F_\beta(p_\epsilon)}[-2 + 2y_1 + \frac{1}{y_1} - \frac{1}{2y_1}] = \frac{6(4y_1^2 - 4y_1 + 1)}{1 - F_\beta(p_\epsilon)}. \end{aligned} \tag{60}$$

Thus, as a piecewise function the PDF is formally,

$$f_{y_1}(y_1) = \begin{cases} \frac{6(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} & \text{for } 0 \leq y_1 \leq \frac{p_\epsilon}{2}, \\ \frac{6(4y_1^2 - 4y_1 + 1)}{1 - F_\beta(p_\epsilon)} & \text{for } \frac{p_\epsilon}{2} \leq y_1 \leq 1/2. \end{cases} \tag{61}$$

Now, the CDF is determined through integration,

$$F_{y_1}(y_1) = \int_0^{y_1} f_{y_1}(y_1) dy_1 = \begin{cases} \frac{6y_1(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} & \text{for } 0 \leq y_1 \leq \frac{p_\epsilon}{2}, \\ \frac{2y_1(4y_1^2 - 6y_1 + 3) + p_\epsilon^2(2p_\epsilon - 3)}{1 - F_\beta(p_\epsilon)} & \text{for } \frac{p_\epsilon}{2} \leq y_1 \leq 1/2. \end{cases} \tag{62}$$

We can integrate the CDF to get,

$$\int_0^{y_1} F_{y_1}(y_1) = \begin{cases} \frac{3y_1^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} & \text{for } 0 \leq y_1 \leq \frac{p_\epsilon}{2}, \\ \frac{8y_1\left(2y_1^3 - 4y_1^2 + 3y_1 + p_\epsilon^2(2p_\epsilon - 3)\right) + p_\epsilon^3(4 - 3p_\epsilon)}{8(1 - F_\beta(p_\epsilon))} & \text{for } \frac{p_\epsilon}{2} \leq y_1 \leq 1/2. \end{cases} \tag{63}$$

Plugging all of this back into Equation 7 yields,

$$\begin{aligned} \hat{b}_i^* &= \begin{cases} p_\epsilon + v_i^p \frac{6v_i^p(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} - \frac{3(v_i^p)^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)}, \\ p_\epsilon + v_i^p \frac{2v_i^p(4(v_i^p)^2 - 6v_i^p + 3) + p_\epsilon^2(2p_\epsilon - 3)}{1 - F_\beta(p_\epsilon)} - \frac{8v_i^p\left(2(v_i^p)^3 - 4(v_i^p)^2 + 3v_i^p + p_\epsilon^2(2p_\epsilon - 3)\right) + p_\epsilon^3(4 - 3p_\epsilon)}{8(1 - F_\beta(p_\epsilon))}, \end{cases} \\ &= \begin{cases} p_\epsilon + \frac{3(v_i^p)^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} & \text{if } 0 \leq v_i^p \leq \frac{p_\epsilon}{2}, \\ p_\epsilon + \frac{8(v_i^p)^2\left(6(v_i^p)^2 - 8v_i^p + 3\right) + p_\epsilon^3(3p_\epsilon - 4)}{8(1 - F_\beta(p_\epsilon))} & \text{if } \frac{p_\epsilon}{2} \leq v_i^p \leq \frac{1}{2}. \end{cases} \end{aligned} \tag{64}$$

Since b_i cannot be larger than 1, we cap the bidding function at one via,

$$b_i^* := \min\{\hat{b}_i^*, 1\}. \tag{65}$$

The utility gained by agent i for using such a bidding function is,

$$u(b_i^*) = \begin{cases} v_i^d - b_i^* + \frac{6(v_i^p)^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} & \text{for } 0 \leq v_i^p \leq \frac{p_\epsilon}{2}, \\ v_i^d - b_i^* + \frac{v_i^p\left(8(v_i^p)^3 - 12(v_i^p)^2 + 6v_i^p + p_\epsilon^2(2p_\epsilon - 3)\right)}{1 - F_\beta(p_\epsilon)} & \text{for } \frac{p_\epsilon}{2} \leq v_i^p \leq 1/2. \end{cases} \tag{66}$$

When this utility is larger than 0, the agent will participate otherwise the agent will not submit a model to the regulator. Finally, we can find the optimal compliance level by using Assumption 2,

$$s_i^* := M^{-1}\big(b_i^*\big). \tag{67}$$

□

D Additional Experiments

Within this section, we verify empirically that our computed PDF and CDFs in Corollaries 1 and 2 are correct. To accomplish this, we randomly sample and compute the product of V_i and λ_i fifty million times. We then plot the PDF and CDF of the resultant products and compare it with our theoretical PDF and CDF. The theoretical PDF and CDF for Corollary 1 are defined in Equations 47 and 48, while those for Corollary 2 are found in Equations 61 and 62. The results of these simulations, which validate our computed PDFs and CDFs, are shown in Figures 6 and 7. To ensure correctness, we perform testing on different values of p_ϵ. As expected, our theory lines up exactly with our empirical simulations for both Corollaries as well as across varying p_ϵ. We note that all experiments are computationally light, with all run locally on an M3 chip with 16GB of RAM.

Finally, we provide the full results of our case study in Section 6 in tabular form below.

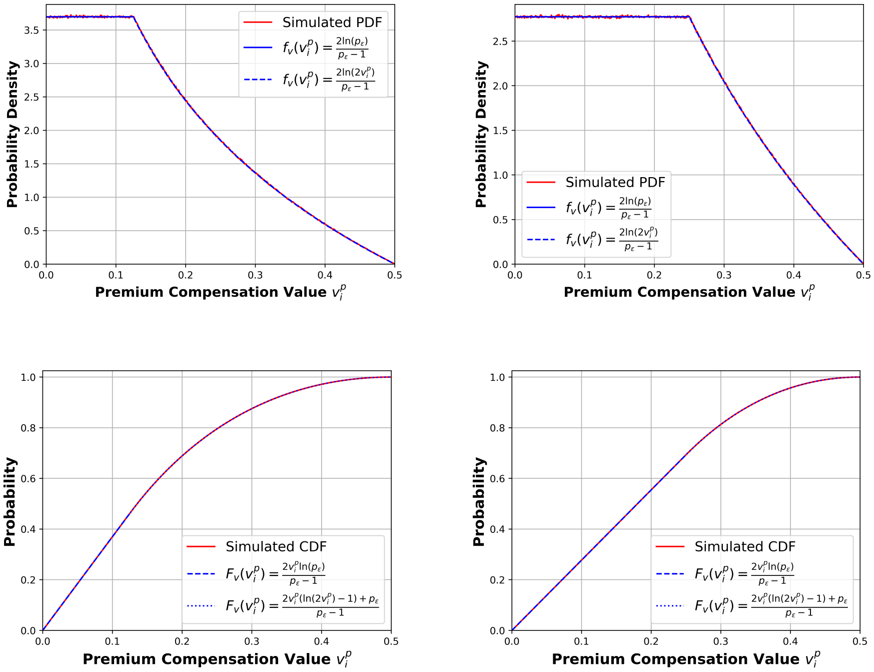


Figure 6: Numerical validation of our derivations for $f_v(v_i^p)$ and $F_v(v_i^p)$, where $v_i^p := V_i\lambda_i$, for V_i and λ_i coming from Uniform distributions (Corollary 1). The price of attaining ϵ is set as $p_\epsilon = 1/4$ (top row) and $p_\epsilon = 1/2$ (bottom row).

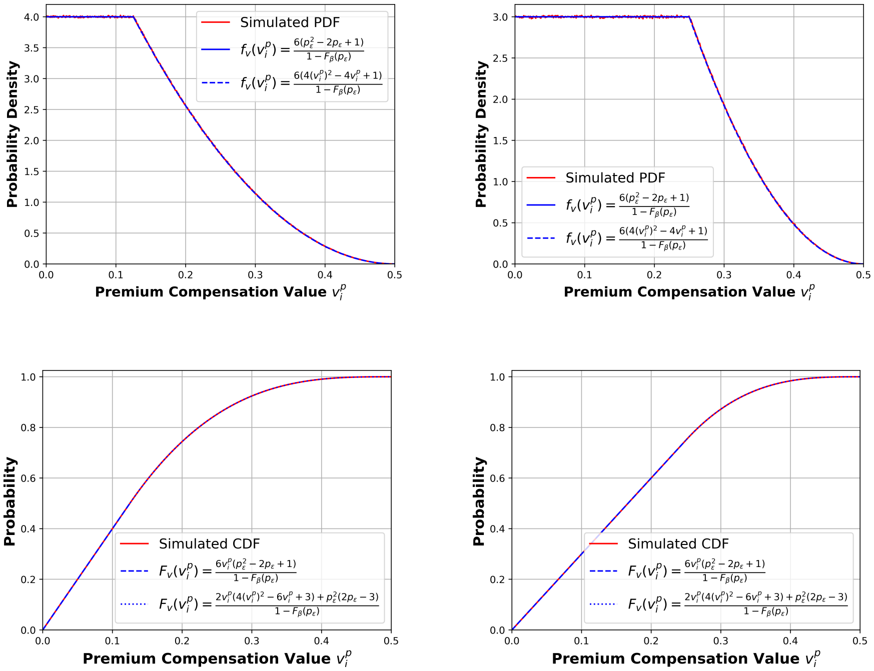


Figure 7: Numerical validation of our derivations for $f_v(v_i^P)$ and $F_v(v_i^P)$, where $v_i^p := V_i\lambda_i$, for V_i coming from a Beta distribution and λ_i from a Uniform distributions (Corollary 2). The price of attaining ϵ is set as $p_\epsilon = 1/4$ (top row) and $p_\epsilon = 1/2$ (bottom row).

Table 3: Equalized Odds as Minority Class Data Increases.

Minority Class %	**Mean Equalized Odds Score**
5%	22.55
10%	22.31
15%	18.97
20%	17.46
25%	15.78
30%	15.44
35%	13.09
40%	11.01
45%	9.83
50%	9.38

E Repeating Circa Auctions

The current auction structure (Algorithm 1) expects agents to submit a single model trained solely for the upcoming auction. There is no expectation that the model will be reused for a future auction, or indication that the model has been submitted to a previous auction. Looking towards the future, we would like to design Circa to fit a repeatable auction structure, in which approved or rejected models may be resubmitted in subsequent auctions.

Repeated Agent Utility. Previously, in Algorithm 1, agents start the regulatory process with zero cost and value (*i.e.,* they are building their models from scratch). In repeating Circa auctions, agent cost and value are accumulated across all previous auction submissions. For example, if an agent trains its already-accepted model further to attain a higher compliance level s_i, its total accumulated training cost is $M(s_i)$. This agent's total value becomes the value its model gained from previous auction submissions plus any value gained from the current auction.

By allowing repeated Circa auctions, an agent is able to repeatedly submit its model for regulatory review. We note that repeated submissions decrease the value of model deployment; once an agent earns the reward for deploying their model, subsequent deployments of the same model with improved compliance levels can be realistically expected to earn less value than the initial deployment. We characterize this loss in value for repeated submissions with an indicator function in the utility function that only allows deployment value to be obtained once, on initial acceptance of a model. While we allow agents to win premium rewards across multiple auctions, we note that a regulator can curb this by either limiting the number of auction submissions per agent or the number of auctions held per year. We now define the repeated Circa auction utility of agent i, who has participated in $a-1$ previous auctions, as:

$$u_{i,a}(b_i) = \left(\sum_{n=1}^{a} \nu_i^n\right) - b_i, \tag{68}$$

where ν_i^n, the value gained at the n^{th} auction model i was submitted to, is formulated as:

$$\nu_i^n = \begin{cases} v_i^{d,n} \cdot 1_{(\text{if } \nu_i^{n-1} = 0)} & \text{if } b_j^n \geq p_\epsilon^n \text{ and } b_i^n < b_j^n \text{ randomly sampled bid } b_j^n, \\ v_i^{d,n} \cdot 1_{(\text{if } \nu_i^{n-1} = 0)} + v_i^{p,n} & \text{if } b_i^n \geq p_\epsilon^n \text{ and } b_i^n > b_j^n \text{ randomly sampled bid } b_j^n, \\ 0 & \text{if n} \leq 0. \end{cases} \tag{69}$$

The repeated Circa auction setup creates a unique property for models in training. If an agent intends to obtain a high compliance level, but an auction takes place mid-training, the agent is actually incentivized to submit their model early if they have a chance at winning the premium reward. Though the model may have a lower likelihood of earning the reward, there is no consequence for models failing to attain the premium reward. Gaining value is strictly beneficial to agents, and accumulated value helps offset the costs of training a model. This property only exists for the premium reward; the deployment reward can only be obtained once, thus there is no incentive to submit early to earn it.

Repeated Optimal Bidding Function. Using the same assumptions for single-auction Circa, namely Assumptions 1 and 2 along with private values, we can derive the bidding function for a rational agent under a repeated Circa auction setting. We follow an equivalent setup to Lemma 1 with regards to the valuation of rewards, giving us the cumulative distribution function for $v_i^p = V_i \lambda_i$ as $F_v(\cdot)$ and the probability distribution function as $f_v(\cdot)$.

From our definition of utility $u_{i,a}(b_i)$, we find that an agent i that does not participate (*i.e.,* submitting $b_i = 0$) receives utility equal to ν_i^a. However, since $b_i = 0$ will never be larger than p_ϵ (by definition), it must be true that $\nu_i^a = 0$ as well, since the model will never meet the required compliance threshold. Therefore, a non-participating agent will always receive non-negative utility.

$$u_{i,a}(0) = 0. \tag{70}$$

Following closely to the proof of Theorem 2 in Appendix C, we find that participating agents $i \in P$ (with P defined in the previous proof) will now have a utility of,

$$\begin{aligned} u_{i,a}(b_i) &= \nu_i^a + v_i^d \cdot 1_{(\nu_i^a = 0)} + v_i^p \mathbb{P}\big(b_i > b_j\big) - b(b_i), \quad b_j \sim \text{randomly sampled agent bid}, \\ &= \nu_i^a + v_i^d \cdot 1_{(\nu_i^a = 0)} + v_i^p F_v(b_i) - b(b_i). \end{aligned} \tag{71}$$

Taking the derivative and setting it equal to zero yields,

$$\frac{d}{db_i} u_{i,a}(b_i) = v_i^p f_v(b_i) - b'(b_i) = 0. \tag{72}$$

As agents bid in proportion to their valuation, we solve the first-order conditions at $b_i = v_i^p$,

$$b'(v_i^p) = v_i^p f_v(v_i^p). \tag{73}$$

Note, at this point in the proof the bidding function calculation is now equivalent to the calculations found in Lemma 1. We can thus follow the same steps to reveal our optimal bidding function,

$$b(v_i^p) := p_\epsilon + v_i^p F_v(v_i^p) - \int_0^{v_i^p} F_v(z) dz, \tag{74}$$

which is equivalent to the optimal bidding function derived in Lemma 1.

As the optimal bidding function is equivalent, calculations for the Nash Bidding Equilibrium are also equivalent to those found in Corollary 1 and Corollary 2. The optimal bid and utility participating in CIRCA Equation 6 under the assumptions of Corollary 1 will thus be,

$$b_i^* := \min\{\hat{b}_i^*, 1\}, \quad \hat{b}_i^* = \begin{cases} p_\epsilon + \frac{(v_i^p)^2 \ln(p_\epsilon)}{p_\epsilon - 1} & \text{if } 0 \le v_i^p \le \frac{p_\epsilon}{2}, \\ p_\epsilon + \frac{8(v_i^p)^2(\ln(2v_i^p) - 1/2) + p_\epsilon^2}{8(p_\epsilon - 1)} & \text{if } \frac{p_\epsilon}{2} \le v_i^p \le \frac{1}{2}, \end{cases}$$

$$u_{i,a}(b_i^*; \boldsymbol{b}_{-i}) = \begin{cases} \nu_i^a + v_i^d \cdot 1_{(\nu_i^a = 0)} + \frac{2(v_i^p)^2 \ln(p_\epsilon)}{p_\epsilon - 1} - b_i^* & \text{if } 0 \le v_i^p \le \frac{p_\epsilon}{2}, \\ \nu_i^a + v_i^d \cdot 1_{(\nu_i^a = 0)} + \frac{2(v_i^p)^2(\ln(2p_\epsilon) - 1) + p_\epsilon}{p_\epsilon - 1} - b_i^* & \text{if } \frac{p_\epsilon}{2} \le v_i^p \le \frac{1}{2}. \end{cases}$$

Agents participating in CIRCA under Corollary 1 submit models with the following compliance,

$$s_i^* := \begin{cases} M^{-1}(b_i^*) > \epsilon & \text{if } u_i(b_i^*; \boldsymbol{b}_{-i}) > 0, \\ 0 \text{ (no model submission)} & \text{else.} \end{cases}$$

The optimal bid and utility participating in CIRCA Equation 6 under the assumptions of Corollary 2 will be,

$$b_i^* := \min\{\hat{b}_i^*, 1\}, \quad \hat{b}_i^* = \begin{cases} p_\epsilon + \frac{3(v_i^p)^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} & \text{if } 0 \le v_i^p \le \frac{p_\epsilon}{2}, \\ p_\epsilon + \frac{8(v_i^p)^2\left(6(v_i^p)^2 - 8v_i^p + 3\right) + p_\epsilon^3(3p_\epsilon - 4)}{8(1 - F_\beta(p_\epsilon))} & \text{if } \frac{p_\epsilon}{2} \le v_i^p \le \frac{1}{2}, \end{cases}$$

$$u_{i,a}(b_i^*; \boldsymbol{b}_{-i}) = \begin{cases} \nu_i^a + v_i^d \cdot 1_{(\nu_i^a = 0)} + \frac{6(v_i^p)^2(p_\epsilon^2 - 2p_\epsilon + 1)}{1 - F_\beta(p_\epsilon)} - b_i^* & \text{for } 0 \le v_i^p \le \frac{p_\epsilon}{2}, \\ \nu_i^a + v_i^d \cdot 1_{(\nu_i^a = 0)} + \frac{v_i^p\left(8(v_i^p)^3 - 12(v_i^p)^2 + 6v_i^p + p_\epsilon^2(2p_\epsilon - 3)\right)}{1 - F_\beta(p_\epsilon)} - b_i^* & \text{for } \frac{p_\epsilon}{2} \le v_i^p \le 1/2. \end{cases}$$

Agents participating in CIRCA under Corollay 2 submit models with the following compliance,

$$s_i^* = \begin{cases} M^{-1}(b_i^*) > \epsilon & \text{if } u_i(b_i^*; \boldsymbol{b}_{-i}) > 0, \\ 0 \text{ (no model submission)} & \text{else.} \end{cases}$$

F FUTURE WORK

While this work addresses key challenges in regulating AI compliance, several directions remain open for future exploration:

(1) Model Evaluation: Creating a realistic protocol for the regulator to evaluate submitted model compliance levels is important to ensure agents do not skirt around compliance requirements. While we leave this problem for future work, one possible solution is that agents can either provide the regulator API access to test its model or provide the model weights directly to the regulator. Truthfulness can be enforced via audits and the threat of legal action.

(2) Extension to Heterogeneous Settings: Extending our mechanism to heterogeneous scenarios, where evaluation data for agents and regulators differs, is a critical next step. Real-world data distributions often vary across contexts, and understanding how these variations affect both model compliance and agent strategies will create a more robust regulatory mechanism. While explicit protocols or mathematical formulations are left as future work, we have a few ideas. One idea could be establishing a data-sharing framework between agents and the regulator, where each participating agent must contribute part of (or all of) its data to the regulator for evaluation. If data can be anonymized, then this would be a suitable solution. Another idea could be that the regulator collects data on its own, and can compare its distribution of data versus each participating agents' data distribution. If distributions greatly differ, then the regulator could collect more data or resort to the previous data-sharing method.