The Backfiring Effect of Weak AI Safety Regulation

Benjamin Laufer[*], Jon Kleinberg[†,§], Hoda Heidari[‡,§]

Abstract

Recent policy proposals aim to improve the safety of general-purpose AI, but there is little understanding of the efficacy of different regulatory approaches to AI safety. We present a strategic model that explores the interactions between safety regulation, the general-purpose AI technology creators, and domain specialists–those who adapt the technology for specific applications. Our analysis examines how different regulatory measures, targeting different parts of the AI development chain, affect the outcome of this game. In particular, we assume AI technology is characterized by two key attributes: *safety* and *performance*. The regulator first sets a minimum safety standard that applies to one or both players, with strict penalties for non-compliance. The general-purpose creator then invests in the technology, establishing its initial safety and performance levels. Next, domain specialists refine the AI for their specific use cases, updating the safety and performance levels and taking the product to market. The resulting revenue is then distributed between the specialist and generalist through a revenue-sharing parameter. Our analysis reveals two key insights: First, weak safety regulation imposed predominantly on domain specialists can backfire. While it might seem logical to regulate AI use cases, our analysis shows that weak regulations targeting domain specialists alone can unintentionally reduce safety. This effect persists across a wide range of settings. Second, in sharp contrast to the previous finding, we observe that stronger, well-placed regulation can in fact mutually benefit *all* players subjected to it. When regulators impose appropriate safety standards on both general-purpose AI creators and domain specialists, the regulation functions as a commitment device, leading to safety and performance gains, surpassing what is achieved under no regulation or regulating one player alone.

1 Introduction

As Generative Artificial Intelligence (AI) and related technologies gain traction, there is an increasing number of proposals for regulation to improve safety. Many of these proposals must at some level grapple with the following question: Who should be targeted with AI

[*] Cornell Tech, bdl56@cornell.edu.
[†] Cornell University.
[‡] Carnegie Mellon University.
[§] Equal contribution.

regulation–the producers of general-purpose AI models[1] or the domain-specialists who adapt the technology for specific use cases? There are seemingly reasonable positions that favor regulating one entity, the other, both, or neither. For example, the downstream domain specialists and deployers are some of the last entities to exert influence on the technology before it interacts with consumers directly, so it is perhaps reasonable that regulation for consumer safety might target requirements at these entities. In contrast, the upstream entities developing general-purpose models exert impact on these models earlier in their development trajectories, facilitating or hindering downstream adoption, which might justify certain regulatory requirements including disclosure mandates [26] and liability standards. Of course, even regulations that solely target one of these actors might impact the other, because their incentives and decisions are intertwined.

We have seen variants of these debates play out as different jurisdictions and policymakers have proposed various regulatory approaches to AI. A number of existing regulation proposals leverage the observation that AI is developed by multiple, interacting actors. Examples include Colorado's AI Act, California's Senate Bill 1047, and the EU AI Act. These frameworks attempt to define the relevant actors, such as base developers and downstream deployers, in order to design conditions and stipulations for determining whether and to whom liability standards, disclosure requirements, or other interventions apply. These conditions and stipulations vary across proposals and policies, with possibly significant implications for the incentives of the players involved in the development of AI technologies and applications.

Modeling the impact of regulatory regimes on AI performance and safety. Given that there are a range of different possible approaches to targeting AI regulation and assessing the impact of each alternative empirically is prohibitive, formal models can enable reasoning about the various regulatory impacts. This paper puts forward a strategic model of the interactions between a general-purpose technology producer (G) and a domain specialist (D), building on the "fine-tuning games" model proposed by Laufer et al. [24]. As the two actors develop an AI technology, they each decide whether and how to invest in two key attributes of technology: *performance*, denoted by α, and *safety*, denoted by β. We assume these actors are operating in a market; each actor experiences some cost for their investment in safety and performance, and obtains a share of the revenue out of the deployment of the AI product/service in the market.

To provide some intuition for what this investment pattern might look like, imagine a firm, G, producing a general-purpose language model that may be used in three domains – say, by healthcare providers (D_1), law firms (D_2), and financial services (D_3). The general-purpose developer moves first, and in light of the particular costs she faces and the anticipated responses from the downstream players, she chooses a certain strategy, represented by a pairing of performance and safety investments (α_0, β_0). Once this investment has been made, the attributes of the technology at this stage can be thought of as akin to a 'base camp,' from which domain specialists may choose to climb further by investing their own effort toward improving the technology's safety and/or performance in their respective domains. Of course, each domain faces their own delicate balance of safety risks and performance

[1] Such AI models are at times referred to as "foundation" or "frontier" models [4, 8]. Throughout this paper, we will use the technology of general-purpose AI to refer to large-scale models that can be adapted to a wide range of tasks and domains.

costs, so the ultimate safety and performance pairs (α_i, β_i) $(i = 1, 2, 3)$ differ across the three domains. See Figure 1 (a) for a visualization of the investment decisions make by G, D_1, D_2, and D_3.

Equipped with this intuition about how these actors behave in an unregulated market, we now turn to our notion of regulation. We conceive of regulation as shaping the game in which players choose their strategies. In particular, this paper will focus specifically on safety regulation. We assume regulation imposes a constraint in the form of a lower bound on the players' choice of safety investment (i.e., β_i's). If a player does not meet the regulatory lower bound on safety, they will be penalized. This regulatory regime can be described using two parameters (θ_G, θ_D), representing the set of thresholds constraining the strategy space of G and D, respectively. The regulation can target the domain-specialist only $(\theta_G = 0, \theta_D > 0)$, the generalist only $(\theta_G = \theta_D > 0)$, both players $(\theta_D > \theta_G > 0)$, or neither $(\theta_G = \theta_D = 0)$. In addition to the decision of who to target, of course, the regulation encodes a decision about what level to set the safety standards. Smaller values of θ are less costly to comply with, and hence capture weaker safety requirements.

First insight: Weak safety regulation can backfire. Turning back to our example in Figure 1, we observe that something striking happens in the second panel, which depicts a scenario where regulation is targeted at the domain-specialist. In this scenario, the safety investment has gotten worse. How could safety regulation – a simple floor dictating a minimum investment level – lead to a less safe product? The mechanism leading to this phenomenon arises because the generalist G is aware of the regulatory safety requirements imposed on domain-specialists, and can use it to her advantage. When the regulator requires that a technology meets a certain level of safety investment by the time it reaches the market, the generalist has an opportunity to engage in a sort of *free-riding* behavior. The generalist is comfortable setting up the base camp at lower altitude, because she knows that the domain-specialist nonetheless has to climb to a level of investment that complies with regulation.[2]

The scenario described above depicts a single instance of a more general phenomenon, which we describe as regulatory *backfiring*. A safety regulation backfires if it yields a total investment in safety lower than the safety investment achieved with no regulation. We identify a number of properties of this phenomenon – for example, backfiring only occurs when the regulation is *weak*, meaning the floor on safety is at or below the level reached in the absence of regulation. Backfiring can occur when D is targeted with regulation or when both G and D are targeted with regulation, but does not occur when only G is targeted. Our results suggest that this non-monotonic effect of regulation occurs for a broad set of games with different cost and revenue functions. Analytically, we prove that backfiring occurs for all quadratic-cost games in which the players invest any non-zero amount in both performance and safety without regulation (these sets of games are defined more formally in Section 4.2).

Second insight: Properly-placed safety regulation can improve the technology and the players' utilities. While weak regulation targeted predominantly at the domain-

[2] Of course, there may be scenarios where the regulation is too high, or the generalist's investment is too low, such that the domain specialist gives up entirely and abstains from any involvement in the technology's production. We handle these scenarios in our analysis of the model, but our point here is to demonstrate that there are circumstances in which regulation gives the generalist leverage to free-ride by lowering the initial investment in safety.

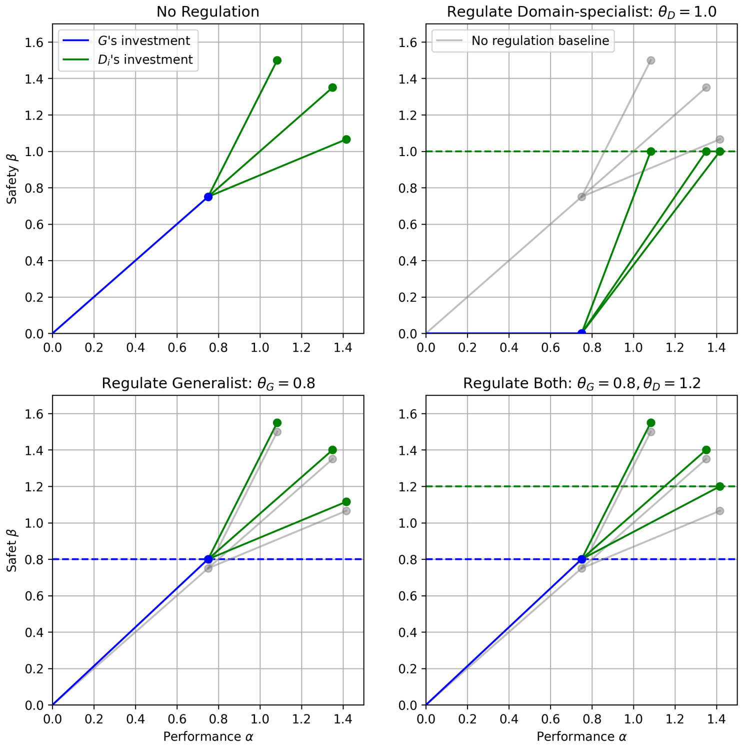

Figure 1: A depiction of a particular instance of our game-theoretic model. This instance of the game consists of one general-purpose producer and three domain-specialists. Each player has a different utility in performance-safety space which dictates the path of development. The no-regulation game (upper left) reveals the players' investment efforts when no floor is imposed on safety. Regulating the domain-specialist alone (upper right) exhibits *backfiring* for all three domains, meaning the regulated safety level is lower than it would be without regulation. In this particular example, the same floor is assumed for all three domain-specialists. Regulating the generalist alone (lower left) improves the safety level slightly across all three domains, compared to no-regulation. Finally, a regime that targets both generalist and specialists with regulation (lower right) is able to 1) retain the improved safety performance from regulating the generalist, 2) improve the safety level of least-safe domain-specialist, while 3) avoiding backfiring.

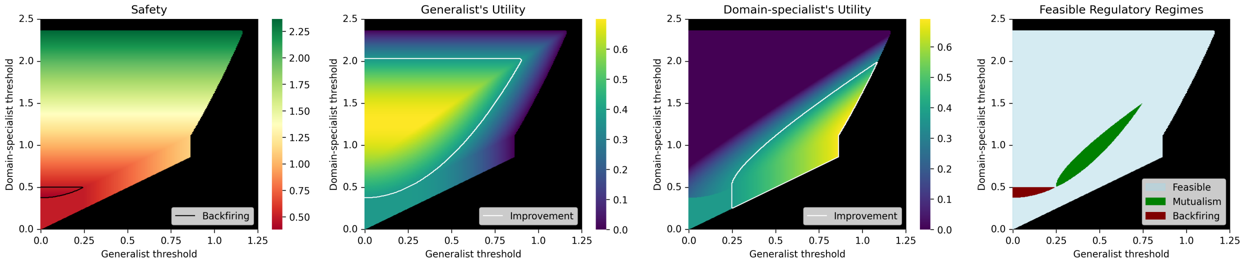

Figure 2: Simulated results for an example of two-player AI regulation model with quadratic costs. Players make costly investments in performance and safety (visualized on left), and then receive some share of revenue that comes from the total investment levels. The players' utilities – defined as their share of the revenue minus the cost of their investment – is visualized for the Generalist (second from left) and the Domain-Specialist (third from left). Colors represent different utility outcomes depending on different combinations of regulatory constraints (θ_G, θ_D) which constrain the players' safety investments. The game is solved over a grid of plausible regulations: $\theta_G \in [0, 1.25]$, $\theta_D \in [\theta_G, 2.5]$ using increments of 0.005, with a total of 105,651 simulated regulation games. Regulations that lead the players to `abstain` are depicted in black. There exists a region where non-zero regulation yields lower safety than no regulation (highlighted on leftmost plot). There also exists a region where regulation yields improvements to each players' utility (highlighted on two center plots). The rightmost plot summarizes our results by showing the backfiring and mutualism outcomes in the θ_G, θ_D space. Parameter values for producing the plot: $C_0 = C_1 = I_2$, $r_\alpha = r_\beta = 1$, and $\delta = 0.5$.

specialist can backfire, our results suggest that other regulatory regimes fare better. When safety standards are directed at both G and D with appropriate strength, regulation can improve not just safety, but the utilities of both players, defined as their revenue share minus their investment cost. This result might seem unintuitive: Regulation only reduces the set of choices available to each actor in our model, so how can regulation lead to choices that mutually benefit both generalist and specialist? What is stopping the players from choosing utility-optimal strategies in the absence of regulation? The reason this phenomenon occurs is a Prisoner's Dilemma-style result: The players' unregulated strategies, which are chosen to maximize their individual utility, fail to yield the strategies that that are globally optimal for both players. By constraining the actors away from the strategies that enable this kind of selfish behavior, regulation can act as a commitment device. The generalist can increase her investments in safety with the assurance that the domain specialist will contribute, too, rather than free-ride off of G's efforts.

Games can exhibit both backfiring and mutualistic regulations, depending on who is targeted and at what threshold. For example, Figure 2 depicts a particular instance of our game setting with one generalist and one domain-specialist. For the particular cost and revenue functions depicted, backfiring regulations and Pareto-improving regulations are possible, and the regulations yielding these effects are visualized. This figure represents a systematic sweep of all pairs of thresholds directed at the generalist, the domain specialist, or both. The pair of thresholds $(0, 0)$ corresponds to the case of no regulation. The safety implications of various regulations are depicted using a red-yellow-green color scale in the leftmost plot, while the utility implications for the generalist and specialist are depicted using a purple-green-yellow color scale in the center plots. Compared to the safety and utility values at the origin points, the backfiring and Pareto-improving regions are regulations which lead to lower safety and higher utilities for both players, respectively. Although this figure depicts an example of a single game, our analysis proves that these backfiring and Pareto-improving regulations exist for a broad class of games with quadratic costs. Namely, we find that backfiring occurs in all games in which the market incentivizes some non-zero investment in both performance and safety without regulation. Our characterization of when this phenomenon occurs includes *separable* scenarios (where the cost of investing in performance is independent of the cost of investing in safety), *complementary* scenarios (where investing in one makes the other cheaper), and weakly *interfering* scenarios (where the cost of investing in one makes the other more expensive) up to a certain bound, which we specify. We provide similar bounds for the mutualism results.

2 Related work

AI Safety Regulation. The rise of AI-related incidents have motivated several AI incident repositories to keep track of common risks [1, 27]. Scholars have attempted to taxonomize AI harms to make sense of the growing array of incidents [32, 37]. Some existing AI risk taxonomies organize risks primarily by *domains.* These include risks to the *physical* or *psychological well-being* of people, *human rights and civil liberties*, *political and economic structures*, *society and culture*, and *the environment* [2]. Others categorize these risks based on how they arise, including *malicious use*, *malfunctions*, or *systemic effects* from wide adoption [5]. In our stylized model, we capture all such considerations using a single scalar that can be toggled by players through investments in safety. Common themes in policy drafts and recommendations stress the importance of balancing the goals of innovation and risk reduction, appropriately defining and targeting thresholds, and the impacts on incentives [12, 18].

Game-theoretic models of AI development. A line of work uses formal models to reason about the strategic and social implications of machine learning (e.g., [7, 14, 21, 22, 25]). More recently, there have been proposals for using modeling approaches to understand the social and safety implications of generative AI [13, 33]. Attempts to model the development process of generative AI often make use of the observation that development is *sequential* and involves *multiple interacting actors* [11]. Many existing works explore different strategic aspects of the market for AI using a stackelberg game. For example, Taitler et al. [35] use a sequential game to explore incentives for data-sharing. Further time-steps, players and decisions have been added to explore particular topics, including the level of openness and market entry dynamics [38, 39]. Taitler and Ben-Porat [34] introduce a particular notion of regulation in a related game-theoretic setting, and similar to our paper, they conceive of regulation as a restriction on the strategy space for developers of generative AI. Though work explicitly examining the interaction between performance and safety attributes in this setting is limited, Jagadeesan et al. [23] explores the interaction between these attributes in a linear regression setting in order to understand firms' market entry decisions.

The fine-tuning games model. Our work builds on and extends the *fine-tuning games* model proposed in Laufer et al. [24]. That model builds a one-dimensional game in which players must bargain over a revenue-sharing contract before investing in performance in sequence. We extend this model in two ways: First, the players' strategy space is two-dimensional in our model, to capture the dynamic that often arises where a regulator wants to steer the technology in a direction (e.g., safety) other than that which is most-profitable (e.g., a baseline combination of performance and safety, dictated by the unregulated market). Second, we introduce the regulation, which can be seen as a *floor* constraining the feasible strategy space of each player. This allows us to explore when targeting generalists, specialists, both or neither is preferable for achieving desiderata like safety.

Economic theory and contracts. Our work leverages pre-existing approaches that are common in the theoretical economics and game theory literatures to reason about the set of possible impacts of AI safety regulation. In particular, we draw inspiration from canonical works in contract theory [20, 31] and the coordination of supply chains [9]. Our model is a variant of a Principal-Agent problem in which the strategy space is defined by two real-valued attributes, and the cost and revenue are functions of these attributes. In this way, our model draws inspiration from Viscusi and Moore [36] analyzing the possible effects of products liability schemes on innovation and safety. That model — a one-player model with no order-of-play effects — demonstrates that liability does not, necessarily, hamper innovation. We assume that innovation is sequential, meaning that an entity's investment in safety or performance builds on the contributions of past investments [6, 19].[3] In what we call the 'no-regulation' game, we assume the players revenue-share via a linear contract [15], a common assumption in the literature (e.g., [3, 10, 16, 17]). However, one way to interpret our mutualism results (Sections 5.3 and 6.2) is as a demonstration that linear contracts are sub-optimal in our setting. Our notion of regulation can be viewed as a set of non-linear contracts defined by a set of strategy constraints, and our results suggest these more expressive contracts can yield higher utility. Of course, still other forms of contracts are possible and may yield different utility implications. We leave these directions to future work.

3 A Model of Regulating AI Safety

Here we offer a formal model for analyzing the effects of regulation on the development of AI applications. Our model is a sequence of sub-games between two players. Each player will choose whether and how to contribute to the technology at a certain point in the development of the technology, and some revenue is received depending on the ultimate attributes of the technology. The players are constrained by regulatory floors on safety, which will be set exogenously by a regulator.

Players. A general-purpose producer, referred to as G, invests in a technology that may be adapted by domain-specialist(s), referred to as D_i. The generalist is the first to invest in the technology, meaning that before G moves, the technology's attributes begin at value 0. Each specialist D_i makes an investment after the generalist has moved.

Technology. We say a technology is described by one or more non-negative attributes

[3] However, some have observed that safety investments can *degrade* as the result of fine-tuning performance investments especially when model weights are open [29, 30]. This scenario, and especially the interaction effects with model openness, are ripe areas for further analysis.

$\gamma \in \mathbb{R}^d$. In this paper, we are interested in two attributes in particular: *performance* and *safety.*[4] Unless otherwise specified, we assume $d = 2$ and that $\gamma = [\alpha, \beta]$ where α refers to performance and β refers to safety.

Economic interests. Each player, acting in a way that maximizes their self-interest, invests some non-zero amount in the technology. G invests to γ_0 and each D_i further invests to γ_i. Accordingly, each must pay a cost for their investment, $\phi_0(\gamma_0)$ and $\phi_i(\gamma_i; \gamma_0)$, respectively. After both players invest, they share a revenue that is brought in as a function of the ultimate attributes of the technology in domain i, $r_i(\gamma_i)$. We assume that, for some $\delta_i \in [0, 1]$, G gets $\delta_i r_i(\gamma_i)$ in revenue and D_i gets $(1 - \delta_i)r_i(\gamma_i)$. δ_i could either be exogenously fixed and given ahead of the game play, or it can be the result of bargaining between G and D_i. When we analyze a game with only one specialist, we will drop the subscript and use δ.

Regulation We model regulation as imposed exogenously on the environment. Regulation is a *minimum constraint* on the safety investment that the players make. A regulation that targets G's investment is characterized by a value $\theta_G \in \mathbb{R}^+$. A non-zero regulation would constrain $G's$ strategy such that $\gamma_0[1] \geq \theta_G$. A regulation targeted at the domain-specialist, similarly, would take the form θ_D and lead the domain-specialist to be constrained in their strategy so $\gamma_i[1] \geq \theta_D$.

Gameplay. The game proceeds as a sequence of subgames:

- Regulation $\{\theta_G, \theta_D\}$ is announced.
- G chooses to either `abstain` or invest in the technology, bringing it to
$$\gamma_0 = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}.$$
- D_i chooses to either `abstain` or invest in the technology, bringing it to
$$\gamma_i = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}.$$
- The technology brings in revenue $r_i(\gamma_i)$, which will be shared such that G receives $\delta_i r_i(\gamma_i)$ and D receives $(1 - \delta_i)r_i(\gamma_i)$.

The utilities of the players are given below:

$$U_G := \sum_i \delta_i r_i(\gamma_i) - \phi_0(\gamma_0); \; U_{D_i} := (1 - \delta_i)r_i(\gamma_i) - \phi_i(\gamma_i; \gamma_0)$$

The best-response sub-game perfect equilibrium strategy for the generalist and specialist, respectively, can be expressed as the following optimization problems:

$$\gamma_0^* := \arg\max_{\gamma_0} U_G \; s.t. \beta_0 \geq \theta_G; \quad \gamma_i^* := \arg\max_{\gamma_i} U_{D_i} \; s.t. \beta_i \geq \theta_G.$$

Finally, the players will opt to `abstain`, if they prefer 0 utility to any other feasible strategy. If either player chooses to `abstain`, then *both* players receive 0 utility.

[4] We note, however, that the results and inferences we draw may hold for other attributes that relate to one another with similar structure.

4 Closed-Form Solutions

In this section, we analyze our regulation game where players' cost functions can be expressed as a two-degree quadratic equation. Specifying a quadratic function over two attributes requires defining a matrix of cost coefficients. The cross-terms in this matrix represent how investments the attributes interact with one another. For the technical portions of the paper, we use the case of one domain specialist (D) as our focus. We therefore have the following cost and revenue functions:

$$\phi_0(\gamma_0) = \gamma_0^T C_0 \gamma_0,$$
$$\phi_1(\gamma_1; \gamma_0) = (\gamma_1 - \gamma_0)^T C_1 (\gamma_1 - \gamma_0),$$
$$r(\gamma_1) = r^T \gamma_1,$$

$$\text{where} \quad C_0 = \begin{bmatrix} c_{0,\alpha\alpha} & c_{0,\alpha\beta} \\ c_{0,\alpha\beta} & c_{0,\beta\beta} \end{bmatrix}; \quad C_1 = \begin{bmatrix} c_{1,\alpha\alpha} & c_{1,\alpha\beta} \\ c_{1,\alpha\beta} & c_{1,\beta\beta} \end{bmatrix}; \quad r = \begin{bmatrix} r_\alpha \\ r_\beta \end{bmatrix}.$$

The players' utilities can thus be expressed as:

$$U_G := \delta r^T \gamma_1 - \gamma_0^T C_0 \gamma_0,$$

$$U_D := (1 - \delta) r^T \gamma_1 - (\gamma_1 - \gamma_0)^T C_1 (\gamma_1 - \gamma_0).$$

It should be noted that not all values for the above parameters correspond to realistic or interesting scenarios. For example, we assume that the diagonal entries of both cost matrices $c_{0,\alpha\alpha}, c_{0,\beta\beta}, c_{1,\alpha\alpha}, c_{1,\beta\beta}$ are non-negative, to capture that investments in goods like safety and performance should have non-zero increasing cost. Although the cross-terms of the cost matrices can be negative, we require that $c_{0,\alpha\beta} > -\sqrt{c_{0,\alpha\alpha} c_{0,\beta\beta}}$ and $c_{1,\alpha\beta} > -\sqrt{c_{1,\alpha\alpha} c_{1,\beta\beta}}$, since it should not be that some combination of investments in α, β come at negative cost. Each players' choices over α and β should be considered as simultaneous across the two attributes, representing a joint optimization over performance and safety.

In this section, we start by providing sub-game perfect equilibria strategies in the case with no regulation, and then provide solutions for the regulated game. The form of problem we are dealing with is a continuous, not-necessarily-convex optimization problem with a constant number of constant-degree polynomials in a constant number of variables. Broadly, the strategy is to put forward a small number of candidate points that must be checked using a limited number of steps. These checks can be implemented numerically. After stating the solved subgame perfect equilibria strategies, we will move to a slate of numerical results and findings analyzing the effects of regulation.

4.1 Subgame perfect equilibria strategies without regulation

In this section, we state the subgame perfect equilibrium strategies to the game under *no regulation*. These can provide intuition about the behaviors in the game, before we add the additional complexity of regulation. These can be seen as a strict generalization of the Fine-Tuning Games solutions [24] to games with two attributes that can interact.

Proposition 4.1. *Given an AI regulation game with quadratic costs, no regulation, and revenue-sharing parameter δ, domain specialist D's subgame perfect equilibrium strategy is one of the values in the following set:*

$$\gamma_1^* \in \left\{ \gamma_0 + \frac{(1-\delta)}{2} C_1^{-1} r, \begin{bmatrix} \alpha_0 \\ \beta_0 + \frac{(1-\delta) r_\beta}{2 c_{1\beta\beta}} \end{bmatrix}, \begin{bmatrix} \alpha_0 + \frac{(1-\delta) r_\alpha}{2 c_{1\alpha\alpha}} \\ \beta_0 \end{bmatrix}, \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} \right\}$$

The strategy is the feasible candidate which maximizes U_D, subject to $U_D \geq 0, \alpha_1 \geq \alpha_0, \beta_1 \geq \beta_0$.

Proposition 4.2. *Given a two-player AI regulation game with quadratic costs, no regulation, and revenue-sharing parameter δ, G's best-response is one of the following candidates:*

$$\gamma_0^* \in \left\{ \frac{\delta}{2} C_0^{-1} r, \begin{bmatrix} 0 \\ \frac{\delta r_\beta}{2 c_{0\beta\beta}} \end{bmatrix}, \begin{bmatrix} \frac{\delta r_\alpha}{2 c_{0\alpha\alpha}} \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

The strategy is the candidate which maximizes U_G, subject to $U_G \geq 0, U_D \geq 0, \alpha_1 \geq 0, \beta_1 \geq 0$.

The proofs of the above two propositions are given in Appendix 8.1. The solutions offer intuition about the set of strategies players might opt to take. They may venture in the direction of some combination of performance and safety, that is, move to a point that does not reside on either constraint. Or, alternatively, they may creep along the axes constraining their strategy space, and invest minimally in either performance or safety.

When do the players prefer one of these strategies over another? In general, our solutions are provided as sets of candidates because there are multiple intersecting constraints that must be checked to ensure a given candidate is optimal. However, our analysis reveals classes of games in which the market will lead players to invest in both safety and performance in conjunction under no regulation. We make this claim formal below.

Remark 4.3. *Given the AI regulation game with quadratic costs, no regulation, and revenue-sharing parameter $\delta \in (0,1)$. If any player p's cost interaction term satisfies the following inequalities:*

$$c_{p,\alpha\beta} < \min\left(\sqrt{c_{p,\alpha\alpha} c_{p,\beta\beta}}, \frac{c_{p,\alpha\alpha} r_\beta}{r_\alpha}, \frac{c_{p,\beta\beta} r_\alpha}{r_\beta} \right),$$

then their best-response strategy includes non-zero investment in both performance and safety.

This claim is proven in Appendix 8.2. The broad intuition is that the first of these inequalities establishes the costs are strictly convex, and the second two ensure that the player's cost interactions are not so positive that investing in both performance and safety is prohibitively expensive compared to investing in one or the other alone. The claim offers some intuition for when a player prefers to invest in both attributes together, even without regulation pushing them to invest in safety. It covers all games in which the cost interactions are negative, which we call the *complementary scenario*, meaning it is cheaper to invest in both performance and safety together than to invest in each individually. It further covers all games in which the cost interactions are zero, which we call the *separable scenario*, meaning there is no benefit or loss to investing in both attributes in conjunction. Finally, it covers certain instances where the cost interactions are positive, which we call the *interfering scenario*, meaning safety investments make performance more costly, and vice versa.

4.2 Subgame perfect equilibria strategies with regulation

Here we provide the subgame perfect equilibria strategies of the two players in our two-attribute game, in the presence of regulation. Notice that the no-regulation gameplay can be derived from these solutions simply by plugging in $\theta_D = \theta_G = 0$. Like the solutions in the prior section, these generalized solutions require checking a number of candidates, but this number has grown to account for the possible responses to regulation.

Proposition 4.4. *Given a two-attribute fine-tuning game with quadratic costs, regulatory constraints θ_G, θ_D, and bargaining parameter δ, the domain specialist D's subgame perfect equilibrium strategy is one of the values in the following set:*

$$\gamma_1^* \in \left\{ \begin{array}{c} \gamma_0 + \tfrac{(1-\delta)}{2} C_1^{-1} r, \begin{bmatrix} \alpha_0 \\ \beta_0 + \tfrac{(1-\delta) r_\beta}{2 c_{1\beta\beta}} \end{bmatrix}, \begin{bmatrix} \alpha_0 + \tfrac{(1-\delta) r_\alpha}{2 c_{1\alpha\alpha}} - \tfrac{c_{1\alpha\beta}}{c_{1\alpha\alpha}} \max(0, \theta_D - \beta_0) \\ \max(\beta_0, \theta_D) \end{bmatrix}, \\ \begin{bmatrix} \alpha_0 \\ \max(\beta_0, \theta_D) \end{bmatrix}, \texttt{abstain}. \end{array} \right\}$$

The strategy is the feasible candidate which maximizes U_D, subject to $U_D \geq 0, \alpha_1 \geq \alpha_0, \beta_1 \geq \max(\beta_0, \theta_D)$.

Proposition 4.5. *Given a two-attribute, two-player fine-tuning game with quadratic costs, regulatory constraints θ_G, θ_D, and bargaining parameter δ, G's best-response is one of the following candidates:*

- $\frac{\delta}{2} C_0^{-1} r$,
- $\begin{bmatrix} 0 \\ \frac{\delta r_\beta}{2 c_{0\beta\beta}} \end{bmatrix}$,
- $\begin{bmatrix} \frac{\delta r_\alpha}{2 c_{0\alpha\alpha}} - \frac{c_{0\alpha\beta}}{c_{0\alpha\alpha}} \theta_G \\ \theta_G \end{bmatrix}$,
- $\begin{bmatrix} 0 \\ \theta_G \end{bmatrix}$,
- $\texttt{abstain}$,
- *Three additional candidates along the $U_D = 0$ constraint, which is given by the following quadratic equation:*

$$(1-\delta) r_\alpha \alpha_0 + \left(\frac{(1-\delta)^2 r_\alpha^2}{4 c_{1\alpha\alpha}} + (1 - r_\beta \theta_D - \frac{c_{1\alpha\beta}}{c_{1\alpha\alpha}} (1-\delta) \theta_D + \frac{c_{1\alpha\beta}^2}{c_{1\alpha\alpha} \theta_D^2} - c_{1\beta\beta} \theta_D^2 \right) + \\ \left(\frac{c_{1\alpha\beta}}{c_{1\alpha\alpha}} (1-\delta) r_\alpha - 2 \frac{c_{1\alpha\beta}^2}{c_{1\alpha\alpha}} \theta_D + 2 c_{1\beta\beta} \theta_D \right) \beta_0 + \left(\frac{c_{1\alpha\beta}^2}{c_{1\alpha\alpha}} - c_{1\beta\beta} \right) \beta_0^2 = 0.$$

The strategy is the candidate which maximizes U_G, subject to $U_G \geq 0, U_D \geq 0, \alpha_1 \geq 0, \beta_1 \geq \theta_G$.

The proof of the above propositions is provided in Appendix 8.3. We outline the intuition behind the proof as follows: Notice that the optimization is an inequality-constrained quadratic optimization problem. The problem has been set up so no solutions exist at infinity, that is, the solutions will either be local maxima or will reside on constraints. Therefore, we can find the critical points for the unconstrained problem, as well as the critical points for every possible combination of every constraint in our problem. This yields a set of candidates, which are worked out and listed in the set above.

There is a bit of additional subtlety in the process for arriving at the last three candidates along the constraint listed at the end of the Proposition. Two of the three candidates reside at the intersection of this constraint with the other constraints—that is, they satisfy the constraint listed and either $\alpha_0 = 0$ or $\beta_0 = \theta_G$. Finding the point that satisfies these combinations of constraints is only as hard as solving the roots of a one-variable quadratic, at worst. The third one, however, is a bit more convoluted. This candidate can be described as the solution to the optimization problem $\max_{\gamma_0} U_G \;\; s.t. U_D = 0$, where the other constraints are ignored. Although this is a (not necessarily convex) quadratic program, specifying the Lagrangian suggests that its solution must be the solution of a system of three distinct equations with three unknown variables $(\alpha_0, \beta_0, \lambda) \in \mathbb{R}^3$. Two of these equations are quadratic, and the other is linear:

- $\delta r_\alpha - 2c_{0\alpha\alpha}\alpha_0 - 2c_{0\alpha\beta}\beta_0 - \lambda(1-\delta)r_\alpha = 0,$
- $\frac{\delta c_{1\alpha\beta} r_\alpha}{c_{1\alpha\alpha}} - 2c_{0\beta\beta}\beta_0 - 2c_{0\alpha\beta}\alpha_0 - \lambda\left(\frac{c_{1\alpha\beta}}{c_{1\alpha\alpha}}(1-\delta)r_\alpha - 2\frac{c_{1\alpha\beta}^2\theta_D}{c_{1\alpha\alpha}} + 2\left(\frac{c_{1\alpha\beta}}{c_{1\alpha\alpha}} - c_{1\beta\beta}\right)\beta_0\right) = 0,$
- The quadratic stated in the proposition.

Though there may be multiple roots satisfying the above equations, the roots are bounded in typical fashion by Bezout's Theorem. Further algebra for arriving at solutions is left to the computer.

5 Computational results

Here we describe a set of numerical tests and demonstrations to explore the strategies in our game, using the solved strategies from the previous section. Our analysis here is focused on the existence of a persistent facet of the model concerning the way the players shift their strategies in response to regulation. With the knowledge that one player or the other is required to meet a regulatory floor, agents can choose their strategies accordingly. In a variety of cases, we observe that the strategies shift in a way that *lowers* the ultimate safety investment compared to safety attained under no regulation. This effect – which we term *backfiring* – is observable in cases where the regulation is weak, meaning it imposes a floor that the players already meet under no regulation.

This section starts by demonstrating the existence of this effect. We then discuss its persistence in cases where players can flexibly choose how they share revenue via a linear

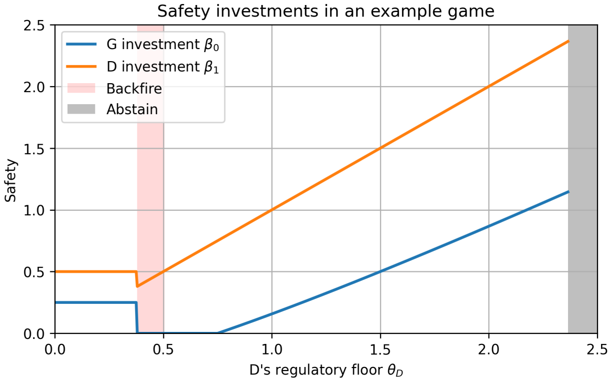

Figure 3: Backfiring observed in a basic two-player game where $\theta_G = 0$ and θ_D is varied over the range $[0, 2.5]$. As θ_D is swept upward from 0, there is some value at which the generalist's score exhibits a discontinuity and the investment in safety lowers. In this example, the discontinuity occurs at a threshold value below the safety attained under no regulation (0.5). In response to this discontinuity in the generalist's strategy, the domain-specialist minimally complies with the regulation, meaning the ultimate safety is reduced for some non-zero regulations. This plot conveys information redundant with the first panel of Figure 2, where the regulation is swept only along the vertical line in which the generalist's threshold is 0.

contract. Finally, in stark contrast to the observation that regulation can backfire, we find that regulation can act as a commitment device, unlocking strategy sequences that mutually benefit the players.

5.1 Regulation can backfire.

Consider a basic game given by the following set of cost and revenue parameters: $C_1 = C_0 = I_2, r_\alpha = r_\beta = 1, \delta = 0.5$. This game is *separable*, meaning there are no interaction effects between performance and safety, and it assumes the market without regulation places equal value on performance and safety. Figure 3 depicts the players' strategies in this game, for varying levels of regulation targeting the Domain-specialist alone. For the lowest regulatory thresholds, we observe that the players stick to their no-regulation safety investments, since they already clear the threshold and their no-regulation investments remain optimal. As the regulatory floor is increased, however, the generalist's strategy exhibits a discontinuity. Crucially, this drop in G's safety investment occurs at a regulatory threshold *lower than* the no-regulation safety strategy.

Why does G switch strategies? Here we attempt to provide some intuition. Abent regulation, there exists some typical best-response that D will take, and G must anticipate this best-response to choose an optimal strategy. Even if G would theoretically prefer D to invest more in safety, G cannot fully control D's actions. Regulation that targets D, however, does just this: it restricts D's strategy space so D must commit to certain safety investments, regardless of G's strategy. Therefore, in the presence of regulation, G is incentivized to select a new strategy sequence with lower investments in safety, because G knows that D will cover

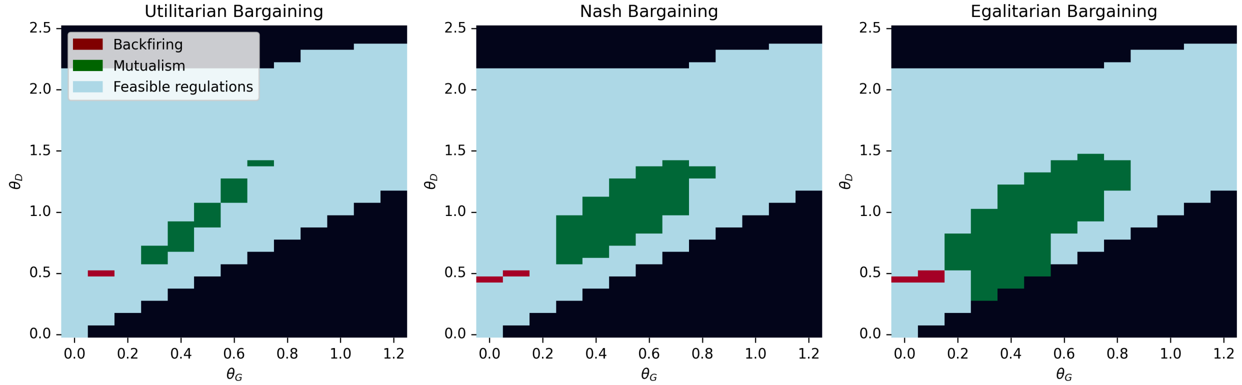

Figure 4: Results from numerical tests over the set of possible (θ_G, θ_D) pairs in the two-attribute, two-player, separable quadratic-cost game. Backfiring occurs in the weak regulatory regimes in which θ_D is just below β_0^A. Regulations that mutually improve both players' utilities over anarchy are detected for all three bargaining solutions. The highest aggregate utility in this game is achieved at $\theta_G = 0.5, \theta_D = 1$.

the gap in safety between G's investment and D's threshold. Put another way, G is given the opportunity to engage in a kind of *free riding* behavior. G creates a gap in the safety investment as a cost-cutting exercise, knowing D must bridge the gap to reap any reward in the game.

Our results here convey that backfiring occurs in one instance of the game. We have yet to give a clear characterization of how widespread this phenomenon is. The example we have shown so far assumes the revenue-sharing parameter is fixed at $\delta = 0.5$. One might imagine that, instead of a fixed revenue sharing arrangement, players can collectively decide how to share the revenue. Does the ability to influence the revenue-sharing parameter prevent cases where weak regulation backfires? We turn to this question next.

5.2 Bargaining does not suffice to prevent backfiring.

Here we provide evidence that the existence of backfiring persists in more games beyond the example portrayed in Figures 2. In particular, we relax the assumption that players share their revenue according to a constant revenue-sharing parameter $\delta = 0.5$. Instead, we allow players to reach *bargaining agreements* to distribute revenue — and, correspondingly, profit — in a way that maximizes their joint utility. Bargaining solutions are arrangements that maximize the players' joint utility.[5] We provide evidence that even when players can distribute revenue in a way that maximizes the joint utility, these arrangements can still exhibit backfiring effects. We assume here that the players jointly agree on a bargaining solution *before* either invests effort, but *after* learning about the regulation.[6] Figure 4 shows the numerical results for a variant of the separable game where we vary the value of δ over 98 values in the range $[0.01, 0.99]$. We vary the regulatory setting for 13 θ_G values in $[0, 1.2]$ and

[5] The relevance of bargaining solutions to our setting is described in further depth by Laufer et al. [24].

[6] The next sections will relax this assumption further, providing findings on the existence of backfiring and mutualism for every non-trivial linear revenue-sharing agreement $\delta \in (0, 1)$.

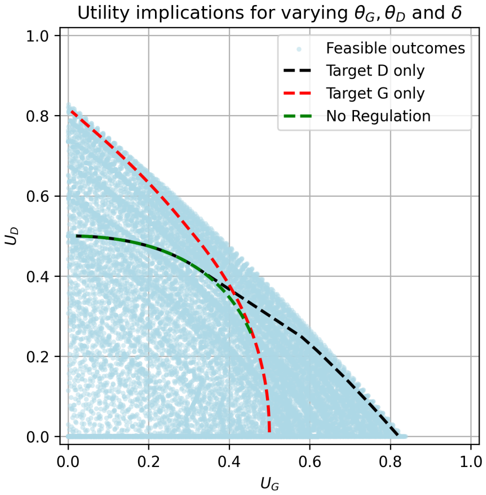

Figure 5: The set of attainable utility outcomes over a grid of possible regulation regimes and bargains for the two-attribute, two-player, separable quadratic-cost game. If we imagine that regulations are *endogenous* to the game – that is, regulations are decided collectively by the players, like δ, then each of the blue points represents a possible game with utility implications for the two players. If the players are restricted to a particular regulatory regime – targeting G only, D only, or neither – then the utility they are able to achieve (depicted in dashed lines) suffers, compared to the regime where both players are subjected to regulation. The shape formed by the light blue dots represents the full set of simulated games, suggesting that regulation targeting both players is at times needed to achieve outcomes that cannot be achieved by regulating just one.

51 θ_D values in $[0, 2.5]$, for a total of 49,686 simulated games. The figure depicts three different processes for arriving at an optimal bargain: *utilitarian*, which selects δ to maximize the sum of utilities, *Nash*, which selects δ to maximize the product of utilities [28], and *egalitarian*, which sets δ to maximize the minimum of the utilities. In all scenarios, we observe at least one instance of a combination of regulations that backfire. Further, we observe a cluster of regulation regimes that yield mutual improvement to utility. The results suggest that even in instances where players can choose how to distribute revenue through revenue-sharing, these agreements are sub-optimal for engendering the right sort of commitment from each of the players, if their goal is to mutually benefit from their interaction. In the next section, we explore the idea that regulation can bring about a *mutualistic* benefit, beyond what is achievable through bargaining.

5.3 Regulation can act as a commitment device.

Here we show that there exist cases where regulation can leave both players *better off* than anarchy, while also benefiting the safety of the technology. Even though the regulation constrains the space of investments that players are able to achieve, it can nonetheless leave each player with higher utility than they are able to achieve under no regulation. To make this finding more clear, we depict the set of all achievable (U_G, U_D) combinations in Figure

5. The light blue cloud of points represents all attainable utility scenarios, over a grid of θ_D, θ_G, and δ values. The dotted lines represent the convex hull (northeastern faces) of attainable utility implications for the following regimes: 1) neither player is targeted with regulation (depicted in green), 2) one player is targeted with regulation (depicted in red and black), and 3) both players are targeted with regulation (inferrable from the outermost feasible points). The figure suggests that a non-vacuous constraint on *both* players achieves more preferable utility outcomes than regulations of individual players or bargaining alone are able to achieve.

These results suggest that, although regulation can backfire, it can also mutually serve the interests of both players while also improving the level of safety of the technology. This finding raises the following question: if it was possible to achieve higher utilities all around, why was this set of strategies not chosen by the players in the unregulated game? The players did not opt for this set of strategies because these strategies are dominated for at least one player in at least one subgame. As a hypothetical, imagine that under no regulation, the players sit down for a conversation before the game, and both say they will contribute ϵ additional investment in safety. When the game reaches the last step, however, D finds he benefits *more* from investing only $\frac{(1-\delta)}{2}C_1^{-1}$, rather than the agreed upon value of $\frac{(1-\delta)}{2}C_1^{-1} + 2\epsilon$. What's more, G *knows* that D will do this, and so G's decision will break the agreement before D even gets the chance to respond. Without the regulation restricting D's behavior away from changing strategies in the final hour, nothing prevents D from pursuing the highest-utility strategy, even if it harms G. Thus, our model has a prisoner's dilemma dynamic baked into it: there are feasible strategies that leave both players better off, but these strategies are not equilibria.

Absent regulation, the players might *wish* they could ensure the other will uphold their side of a verbal agreement, though they are unable to guarantee it. Regulation, therefore, can act as a *commitment device*, which lends teeth to agreements that the players are able to enter prior to making their investments. This commitment device can be valuable in a formal sense: Both players would be willing to pay for it, as long as the price is less than the amount of utility they collectively gain under regulation.

6 A General Characterization

In the previous sections, we arrived at closed-form solutions for the players' strategies and have demonstrated individual instances that exhibit the backfiring effect of regulation. We have not yet determined how widespread this phenomenon is. In this section, we provide analytical results that characterize when this phenomenon occurs. Our findings suggest that this effect is notably widespread. We find that for all quadratic-cost games, backfiring occurs as long as both of the technology's attributes (performance and safety) are sufficiently *complementary* such that, under no regulation, the players will invest in some combination of them. Intuitively, if the players invested only in performance under no regulation, backfiring would be impossible as the baseline safety investment would be zero. Therefore, our condition for backfiring covers all games where the market prefers some non-zero baseline investment in performance and safety. The condition we rely on is precisely the condition introduced in Remark 4.3, which represents an upper bound on the cost interaction terms. This section will prove that both backfiring and mutualism occur in a range of scenarios that depend crucially on the cost interaction term, and will describe what this dependence looks like.

6.1 Backfiring occurs in all mixed-strategy games.

Below we prove that for all AI regulation games in which the players invest a non-zero amount in safety and performance under no regulation, there is a non-empty set of regulatory regimes that exhibit a backfiring effect.

Theorem 6.1. *Given an AI regulation game with quadratic costs. If both players' cost interactions meet the following conditions:*

$$c_{p,\alpha\beta} < \min\left(\sqrt{c_{p,\alpha\alpha}c_{p,\beta\beta}}, \frac{c_{p,\alpha\alpha}r_\beta}{r_\alpha}, \frac{c_{p,\beta\beta}r_\alpha}{r_\beta}\right),$$

then there exists an $\epsilon > 0$ such that the regulatory regime $\theta_G = 0, \theta_D = \beta_0^A - \epsilon$ backfires.

The proof of the above theorem is provided in Appendix 10. Here we provide an overview of the conceptual argument. We start by observing that the unregulated optimal strategies γ_0^A, γ_1^A remain feasible in weak regulatory settings. These strategies dominate all alternative strategies in which the players contribute to safety *beyond* their regulatory constraints, as any such strategy was available in the no regulation scenario, so they were already shown to be sub-optimal compared to γ_0^A, γ_1^A. The proof's task, therefore, is is to find some $\theta_D < \beta_1^A$ and some $\gamma_0' \neq \gamma_0^A$, such that D *minimally complies* with the regulation ($\beta_0' = \theta_D$), and further, $U_G(\gamma_0'; \theta_D) > U_G(\gamma_0^A; \theta_D)$. For the proof to work, we choose a regulation of $\theta_G = 0, \theta_D = \beta_0^A - \epsilon$ for some small positive $\epsilon > 0$, and generalist strategy $\gamma_0' = \begin{bmatrix} \frac{\delta r_\alpha}{2c_{0,\alpha\alpha}}\left(\beta_0^A - 2\epsilon\right) \\ \beta_0^A - 2\epsilon \end{bmatrix}$. For sufficiently small ϵ, we find that the change to the utility of G for using this strategy is positive as long as the following condition is met: $r_\beta > \frac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}} r_\alpha$. This inequality, given by the analysis in Appendix 10, is precisely the condition established in Remark 4.3 for non-zero investment in safety under no regulation.

The above results demonstrate that backfiring does not only exist in single degenerate cases: It occurs in a range of scenarios in which players share revenue and each contribute non-zero effort to the development of the technology. These scenarios include settings in which the two attributes are *complementary*, as well as a range of settings where the two attributes are *interfering*, up to a particular limit that we are able to specify. We note that further generalizations are open for broader functional forms, including more expressive polynomial costs and exponential costs. The generality of the backfiring effect in the quadratic case gives us reason to believe that the effect might hold for a broader set of forms, though we leave these directions to future work.

6.2 Mutualism occurs in sufficiently separable games.

So far, we have shown that a set of regulations backfire in a swath of two-attribute games. Here we provide a second result on a set of regulations that fare better. Using similar logic about games with bounded interaction effects between the the two attributes, we find that there exist combinations of regulatory thresholds that mutually improve the two players' utilities, as well as the safety level of the technology. We state this result below.

Theorem 6.2. *Given a two-player AI regulation game with quadratic costs. If both players meet the following conditions:*

$$|c_{p,\alpha\beta}| < \min\left(\sqrt{c_{p,\alpha\alpha}c_{p,\beta\beta}}, \frac{c_{p,\alpha\alpha}r_\beta}{r_\alpha}, \frac{c_{p,\beta\beta}r_\alpha}{r_\beta}\right),$$

then there exists an $\epsilon > 0$ such that the regulatory regime $\theta_G = \beta_0^A + \epsilon, \theta_D = \beta_1^A + 2\epsilon$ mutually improves both players' utilities.

The proof of the above theorem is given in Appendix 11. The proof follows a similar strategy to the backfiring proof. We are focused on the set of games where the players arrive at unconstrained solutions in the case of no regulation, and we perturb the regulation by a small positive ϵ value and see the implications for the players' utilities. Here, instead of targeting only the domain-specialist and specifying a threshold slightly below the unconstrained optimal strategy, we set the regulation to target *both players* using a threshold slightly *above* their unconstrained strategies. Instead of measuring the impact on safety, we measure the impact on the players' utilities and find that, under the specified condition, the utilities both improve.

The results suggest that, similar to the characterization of backfiring, the mutualism effect is observable in a range of quadratic-cost games, including in *separable scenarios* and a range of *complementary* and *interfering* scenarios. Notice, however, that our condition for establishing when mutualism occurs is slightly different than the condition in the backfiring theorem. Instead of a one-sided bound on the players' cost interaction terms, our proof relies on a two-sided bound. The analysis suggests there may be certain games where the two attributes are *strongly complementary* where slightly increasing the regulation in the manner proposed does not increase players' utilities. In other words, if the market already sufficiently incentivizes joint investments in safety and performance, then forcing safety requirements on both players in equal proportion may not benefit players' utilities. In these cases, a linear contract may suffice to serve the utilities of the players, and so regulation would only be needed for achieving the goal of advancing safety, and would not serve the additional role in enforcing commitments from players.

7 Conclusion

Proposals for AI regulation have made use of the idea that different entities contribute to these technologies in succession. This work provides a model for reasoning about the effects of targeting AI safety regulation along the development chain. Our findings suggest that *weak* safety regulation predominantly targeted at the domain specialist can backfire, yielding lower investments in safety than in the alternative case of no regulation. Our findings further suggest that regulation appropriately targeted at both upstream producers and downstream specialists can exhibit a mutualism effect in which both entities benefit. After demonstrating instances of the backfiring and mutualism effects through a numerical simulation, we provide analysis showing these phenomena are not just degenerate cases but hold in a range of scenarios.

Our results reveal natural directions for future research. In the setting we have put forward, it would be interesting to move beyond showing the existence of backfiring and mutualism regions and characterize the shape of these regions and the magnitude of their effects. Certain segments of the boundaries of these regions are straightforward but others seem to require solving higher-order polynomials to express in closed-form.

Generalizations beyond the quadratic-cost games might be interesting. For instance, it may be possible to show that backfiring and Pareto-improvement effects occur for any convex cost and concave revenue games meeting where there exist some marginal conditions on the functions' marginal conditions including their slopes and intercepts.

We have predominantly focused on the case where there is one domain-specialist, but in many real-world settings the development of AI technologies involve multiple domains, and each domain may involve many entities who compete. To what extent does competition between multiple entities change the backfiring and Pareto-improving impacts of regulation? Pursuing questions about multiple domain-specialists would require further specifying the structure of G's contract with each specialist, which might reasonably be conceived as a constant revenue share across domains, a constant fixed price across domains, or a variable price across domains. Relatedly, approaches to regulating different specialists may be conceived of as domain-specific (different requirements for each domain) or domain-agnostic (requirements for all domains). Pursuing questions about multiple generalists may also illuminate interesting directions. In particular, if different domains have different preferences over attributes, there may be scenarios where general providers *specialize* their investments to capture some domains and cede others to their competitors. Such dynamics raise new questions about how to design regulation to account for these rich constellations of interacting actors.

Acknowledgements.

The authors thank the AI, Policy and Practice (AIPP) group at Cornell University and the Digital Life Initiative (DLI) at Cornell Tech. We thank James Grimmelmann, Aleksandra Korolova, Karen Levy, Hamidah Orderinwale, Helen Nissenbaum, Tori Qiu, and Manish Raghavan for helpful conversations and comments.

This work is supported by a grant from the John D. and Catherine T. MacArthur Foundation. Ben Laufer is also supported by a doctoral fellowship from DLI, a LinkedIn-Bowers CIS PhD Fellowship, and a SaTC NSF grant CNA-1704527. Jon Kleinberg is also supported by a Vannevar Bush Faculty Fellowship, AFOSR aware FA9550-19-1-0183, and a grant from the Simons Foundation. Hoda Heidari acknowledges support from NSF (IIS-2040929 and IIS-2229881) and PwC (through the Digital Transformation and Innovation Center at CMU). Any opinions, findings, conclusions, or recommendations expressed in this material do not reflect the views of NSF or other funding agencies.

References

[1] Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, Mark A. Sayre, Ushnish Sengupta, Arthit Suriyawongkul, Ruby Thelot, Sofia Vei, and Laura Waltersdorfer. AIAAIC Harms taxonomy - v1.7.pdf. URL `https://drive.google.com/file/u/1/d/1-dqNDv2G43856tABKP-CSYTyAL8xo2kj/view?usp=sharing&usp=embed_facebook`.

[2] Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, Mark A. Sayre, Ushnish Sengupta, Arthit Suriyawongkul, Ruby Thelot, Sofia Vei, and Laura Waltersdorfer. A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms, July 2024. URL `http://arxiv.org/abs/2407.01294`. arXiv:2407.01294 [cs].

[3] Tal Alon, Paul Dütting, Yingkai Li, and Inbal Talgam-Cohen. Approximate optimality of linear contracts under uncertainty. *arXiv preprint arXiv:2211.06850*, 2022.

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

[5] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. International ai safety report. *arXiv preprint arXiv:2501.17805*, 2025.

[6] James Bessen and Eric Maskin. Sequential innovation, patents, and imitation. *The RAND Journal of Economics*, 40(4):611–635, 2009.

[7] Avrim Blum, Nika Haghtalab, Richard Lanas Phillips, and Han Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *International Conference on Machine Learning*, pages 1005–1014. PMLR, 2021.

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[9] Gérard P Cachon. Supply chain coordination with contracts. *Handbooks in operations research and management science*, 11:227–339, 2003.

[10] Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105(2): 536–563, 2015.

[11] Sarah Huiyi Cen, Aspen Hopkins, Andrew Ilyas, Aleksander Madry, Isabella Struckman, and Luis Videgaray Caso. Ai supply chains. 2023.

[12] Jennifer Tour Chayes, Mariano-Florentino Cuéllar, and Li Fei-Fei. Draft report of the joint california policy working group on ai frontier models. *Draft, Joint California Policy Working Group on AI Frontier Models*, 2025.

[13] Sarah Dean, Evan Dong, Meena Jagadeesan, and Liu Leqi. Accounting for ai and users shaping one another: The role of mathematical models. *arXiv preprint arXiv:2404.12366*, 2024.

[14] Kate Donahue and Jon Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5303–5311, 2021.

[15] Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 369–387, 2019.

[16] Paul Dütting, Tomer Ezra, Michal Feldman, and Thomas Kesselheim. Multi-agent contracts. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1311–1324, 2023.

[17] Paul Dütting, Tomer Ezra, Michal Feldman, and Thomas Kesselheim. Multi-agent combinatorial contracts. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1857–1891. SIAM, 2025.

[18] Matthew R Gaske. Regulation priorities for artificial intelligence foundation models. *Vand. J. Ent. & Tech. L.*, 26:1, 2023.

[19] Jerry R Green and Suzanne Scotchmer. On the division of profit in sequential innovation. *The Rand journal of economics*, pages 20–33, 1995.

[20] Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. In *Foundations of insurance economics: Readings in economics and finance*, pages 302–340. Springer, 1992.

[21] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[22] Keegan Harris, Hoda Heidari, and Steven Z Wu. Stateful strategic regression. *Advances in Neural Information Processing Systems*, 34:28728–28741, 2021.

[23] 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.

[24] Benjamin Laufer, Jon Kleinberg, and Hoda Heidari. Fine-tuning games: Bargaining and adaptation for general-purpose models. In *Proceedings of the ACM on Web Conference 2024*, pages 66–76, 2024.

[25] Lydia T Liu, Nikhil Garg, and Christian Borgs. Strategic ranking. In *International Conference on Artificial Intelligence and Statistics*, pages 2489–2518. PMLR, 2022.

[26] Shayne Longpre, Kevin Klyman, Ruth E Appel, Sayash Kapoor, Rishi Bommasani, Michelle Sahar, Sean McGregor, Avijit Ghosh, Borhane Blili-Hamelin, Nathan Butters, et al. In-house evaluation is not enough: Towards robust third-party flaw disclosure for general-purpose ai. *arXiv preprint arXiv:2503.16861*, 2025.

[27] Sean McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database. URL `https://incidentdatabase.ai/`.

[28] John F Nash et al. The bargaining problem. *Econometrica*, 18(2):155–162, 1950.

[29] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

[30] Xiangyu Qi, Boyi Wei, Nicholas Carlini, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, and Peter Henderson. On evaluating the durability of safeguards for open-weight llms. *arXiv preprint arXiv:2412.07097*, 2024.

[31] Stephen A Ross. The economic theory of agency: The principal's problem. *The American economic review*, 63(2):134–139, 1973.

[32] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction, July 2023. URL `http://arxiv.org/abs/2210.05791`. arXiv:2210.05791 [cs].

[33] Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. Game theory meets large language models: A systematic survey. *arXiv preprint arXiv:2502.09053*, 2025.

[34] Boaz Taitler and Omer Ben-Porat. Selective response strategies for genai. *arXiv preprint arXiv:2502.00729*, 2025.

[35] Boaz Taitler, Omer Madmon, Moshe Tennenholtz, and Omer Ben-Porat. Data sharing with a generative ai competitor. *arXiv preprint arXiv:2505.12386*, 2025.

[36] W Kip Viscusi and Michael J Moore. Product liability, research and development, and innovation. *Journal of Political Economy*, 101(1):161–184, 1993.

[37] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, Amelia Glaese, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York,

NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533088. URL `https://dl.acm.org/doi/10.1145/3531146.3533088`.

[38] Yanxuan Wu, Haihan Duan, Xitong Li, and Xiping Hu. Navigating the deployment dilemma and innovation paradox: Open-source versus closed-source models. In *Proceedings of the ACM on Web Conference 2025*, pages 1488–1501, 2025.

[39] Fasheng Xu, Xiaoyu Wang, Wei Chen, and Karen Xie. The economics of ai foundation models: Openness, competition, and governance. *Competition, and Governance (August 11, 2024)*, 2024.

8 Game Solving

8.1 Player's strategies without regulation

The domain-specialist's strategy. The proof for Proposition 4.1 is given below.

Proof. D's best-response strategy is the value γ_1^* that maximizes D's utility.

$$\gamma_1^*(\gamma_0, \delta) \;=\; \arg\max_{\gamma_1} U_D(\gamma_0, \gamma_1 \delta) \text{ s.t. } U_D \geq 0, \alpha_1 \geq \alpha_0, \beta_1 \geq \beta_0$$

Observe that D will not abstain because zero-investment ($\gamma_1 = \gamma_0$) is cost-free, yielding non-negative utility, so we can safely ignore the constraint. To solve the optimization, we specify the Lagrangian as follows for some multipliers $\lambda_1 \in \mathbb{R}, \lambda_2 \in \mathbb{R}$ and a slack variables $s_1 \in \mathbb{R}, s_2 \in \mathbb{R}$. By construction, we assert that the slack variables are only non-zero when the multipliers are zero, and the multipliers are non-zero only if the slack variables are zero.

$$\mathcal{L} := (1-\delta) r^T \gamma_1 - (\gamma_1 - \gamma_0)^T C_1 (\gamma_1 - \gamma_0) - \lambda_1(\alpha_1 - \alpha_0 - s_1^2) - \lambda_2(\alpha_1 - \alpha_0 - s_2^2).$$

We partially differentiate with respect to each decision variable and each multiplier.

$$\begin{aligned}
\frac{\partial}{\partial \alpha_1}\mathcal{L} &= 0 \\
\iff (1-\delta) r_\alpha - 2c_{1,\alpha\alpha}(\alpha_1 - \alpha_0) + 2c_{1,\alpha\beta}(\beta_1 - \beta_0) - \lambda_1 &= 0 \\
\frac{\partial}{\partial \beta_1}\mathcal{L} &= 0 \\
\iff (1-\delta) r_\beta - 2c_{1,\beta\beta}(\beta_1 - \beta_0) + 2c_{1,\alpha\beta}(\alpha_1 - \alpha_0) - \lambda_2 &= 0 \\
\frac{\partial}{\partial \lambda_1}\mathcal{L} &= 0 \\
\iff -\alpha_1 + \alpha_0 + s_1^2 &= 0 \\
\frac{\partial}{\partial \lambda_2}\mathcal{L} &= 0 \\
\iff -\beta_1 + \beta_0 + s_2^2 &= 0
\end{aligned}$$

Using complementary slackness, we have four possible options:

1. $s_1 = 0, \lambda_1 > 0, s_2 = 0, \lambda_2 > 0 \to \beta_1^* = \beta_0, \alpha_1^* = \alpha_0$.
2. $s_1 \neq 0, \lambda_1 = 0, s_2 = 0, \lambda_2 > 0 \to \beta_1^* = \beta_0$, and we can plug into our first of four equations above:

$$\begin{aligned} (1-\delta)r_\alpha - 2c_{1,\alpha\alpha}(\alpha_1 - \alpha_0) + 2c_{1,\alpha\beta}(\beta_1 - \beta_0) - \lambda_1 &= 0 \\ \to (1-\delta)r_\alpha - 2c_{1,\alpha\alpha}(\alpha_1 - \alpha_0) &= 0 \\ \to \alpha_1^* = \alpha_0 + \frac{(1-\delta)r_\alpha}{2c_{1,\alpha\alpha}}. \end{aligned}$$

3. $s_1 = 0, \lambda_1 > 0, s_2 \neq 0, \lambda_2 = 0 \to \alpha_1^* = \alpha_0$, and we can plug in to equation 2:

$$\begin{aligned} (1-\delta)r_\beta - 2c_{1,\beta\beta}(\beta_1 - \beta_0) + 2c_{1,\alpha\beta}(\alpha_1 - \alpha_0) - \lambda_2 &= 0 \\ \to (1-\delta)r_\beta - 2c_{1,\beta\beta}(\beta_1 - \beta_0) - \lambda_2 &= 0 \\ \to \beta_1^* = \beta_0 + \frac{(1-\delta)r_\beta}{2c_{1,\beta\beta}}. \end{aligned}$$

4. $s_1 \neq 0, \lambda_1 = 0, s_2 \neq 0, \lambda_2 = 0 \to$ This is the unconstrained critical point, and is solved via the first two systems of equations:

$$\begin{aligned} \nabla U_D = (1-\delta)r - 2C_1(\gamma_1 - \gamma_0) &= 0 \\ \to \gamma_1^* = \gamma_0 + \frac{(1-\delta)}{2}C_1^{-1}r. \end{aligned}$$

Thus we have established our four candidates in the proposition statement. □

The Generalist's strategy. The proof for Proposition 4.2 is given below.

Proof. G's best-response strategy is the value γ_0^* that maximizes G's utility.

$$\gamma_0^*(\delta) \quad = \quad \arg\max_{\gamma_1} U_G(\gamma_0, \delta) \text{ s.t. } U_G \geq 0, \alpha_0 \geq 0, \beta_0 \geq 0.$$

Following the same steps as the proof of Proposition 4.1, we specify the Lagrangian as follows for multipliers $\lambda_1 \in \mathbb{R}, \lambda_2 \in \mathbb{R}$ and a slack variables $s_1 \in \mathbb{R}, s_2 \in \mathbb{R}$.

$$\mathcal{L} := \delta r^T \gamma_1 - \gamma_0^T C_0 \gamma_0 - \lambda_1(\alpha_0 - s_1^2) - \lambda_2(\alpha_0 - s_2^2).$$

We partially differentiate with respect to each decision variable and each multiplier.

$$\begin{aligned} \frac{\partial}{\partial \alpha_0}\mathcal{L} &= 0 \\ \iff \delta r_\alpha - 2c_{0,\alpha\alpha}\alpha_0 + 2c_{0,\alpha\beta}\beta_0 - \lambda_1 &= 0, \\ \frac{\partial}{\partial \beta_1}\mathcal{L} &= 0 \\ \iff \delta r_\beta - 2c_{0,\beta\beta}\beta_0 + 2c_{1,\alpha\beta}\alpha_0 - \lambda_2 &= 0, \\ \frac{\partial}{\partial \lambda_1}\mathcal{L} &= 0 \\ \iff -\alpha_0 + s_1^2 &= 0, \\ \frac{\partial}{\partial \lambda_2}\mathcal{L} &= 0 \\ \iff -\beta_0 + s_2^2 &= 0. \end{aligned}$$

Using complementary slackness, we have four possible options:

1. $s_1 = 0, \lambda_1 > 0, s_2 = 0, \lambda_2 > 0 \to \beta_0^* = 0, \alpha_0^* = 0.$

2. $s_1 \neq 0, \lambda_1 = 0, s_2 = 0, \lambda_2 > 0 \to \beta_0^* = 0$, and we can plug into our first of four equations above:

$$\begin{aligned} \delta r_\alpha - 2c_{0,\alpha\alpha}\alpha_0 + 2c_{0,\alpha\beta}\beta_0 - \lambda_1 &= 0 \\ \to \delta r_\alpha - 2c_{0,\alpha\alpha}\alpha_0 &= 0 \\ \to \alpha_0^* = \frac{\delta r_\alpha}{2c_{0,\alpha\alpha}}. \end{aligned}$$

3. $s_1 = 0, \lambda_1 > 0, s_2 \neq 0, \lambda_2 = 0 \to \alpha_0^* = 0$, and we can plug in to equation 2:

$$\begin{aligned} \delta r_\beta - 2c_{0,\beta\beta}\beta_0 + 2c_{0,\alpha\beta}\alpha_0 - \lambda_2 &= 0 \\ \delta r_\beta - 2c_{0,\beta\beta}\beta_0 &= 0 \\ \to \beta_0^* = \frac{\delta r_\beta}{2c_{0,\beta\beta}}. \end{aligned}$$

4. $s_1 \neq 0, \lambda_1 = 0, s_2 \neq 0, \lambda_2 = 0 \to$ This is the unconstrained critical point, and is solved via the first two systems of equations:

$$\begin{aligned} \nabla U_G = \delta r - 2C_0\gamma_0 &= 0 \\ \to \gamma_0^* = \frac{\delta}{2}C_0^{-1}r. \end{aligned}$$

Thus we have established our four candidates. □

8.2 Condition for non-zero performance and safety investment

Condition establishing non-zero investment. Below we prove Remark 4.3.

Proof. The first of the three inequalities establishes that the player's costs are strictly convex:

$$c_{\alpha\beta} < \sqrt{c_{p,\alpha\alpha}c_{p,\beta\beta}} \iff c_{p,\alpha\alpha}c_{p,\beta\beta} - c_{\alpha\beta}^2 > 0 \iff \det C_p > 0.$$

By the spectral theorem, we know a 2x2 matrix is positive definite if and only if its determinant and trace are both positive, which is now established. By Lemma 9.1, the utility is strictly concave for our setting if and only if the cost is strictly convex. Thus the unconstrained solution is the global optimum as long as it is feasible. Thus, the necessary and sufficient condition for optimality is the condition for feasibility.

- For the generalist:

  $$\frac{\delta}{2}C_0^{-1}r > 0 \iff \frac{\delta}{2\det C_0}\begin{bmatrix} c_{0,\beta\beta}r_\alpha - c_{0,\alpha\beta}r_\beta \\ -c_{0,\alpha\beta}r_\alpha + c_{0,\alpha\alpha}r_\beta \end{bmatrix} > \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

  Using the same positive definiteness identity above, we know the determinant is positive. We are given $\delta > 0$. Thus we can cancel the positive constant term $\frac{\delta}{2\det C_0}$. The two inequalities simplify to those stated in the proposition.

- For the specialist, the proof proceeds identically. Observe that $(1-\delta) \geq 0$ and the unconstrained contribution is given by: $\frac{1-\delta}{2}C_1^{-1}r$.

□

8.3 Proof for Player Strategies with Regulation

Here we provide proofs for our propositions establishing best-response strategies for the players.

Domain-specialist best-response under regulation. Here we provide the proof of Proposition 4.4, the domain specialist's best response under regulatory requirement θ_D.

Proof. D's best-response strategy is the value γ_1^* that maximizes D's utility. D will abstain if and only if the best option yields negative utility.

$$\gamma_1^*(\gamma_0, \delta, \theta_D) = \arg\max_{\gamma_1} U_D(\gamma_0, \delta, \theta_D) \text{ s.t. } U_D \geq 0, \alpha_1 \geq \alpha_0, \beta_1 \geq \max(\beta_0, \theta_D).$$

Define $\kappa = \max(\beta_0, \theta_D)$. To solve the optimization, we specify the Lagrangian as follows for some multipliers $\lambda \in \mathbb{R}^3$ and a slack variables $s \in \mathbb{R}^3$.

$$\mathcal{L} := (1-\delta)r^T\gamma_1 - (\gamma_1 - \gamma_0)^T C_1(\gamma_1 - \gamma_0) - \lambda_1(\alpha_1 - \alpha_0 - s_1^2) - \lambda_2(\beta_1 - \kappa - s_2^2) - \lambda_3(U_D - s_3^2).$$

We partially differentiate with respect to each decision variable and each multiplier.

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_1}\mathcal{L} &= 0 \\
\iff (1-\delta)r_\alpha - 2c_{1,\alpha\alpha}(\alpha_1-\alpha_0) + 2c_{1,\alpha\beta}(\beta_1-\kappa) - \lambda_1 - \lambda_3\frac{\partial U_D}{\partial \alpha_1} &= 0 \\
\iff (1-\lambda_3)\left((1-\delta)r_\alpha - 2c_{1,\alpha\alpha}(\alpha_1-\alpha_0) + 2c_{1,\alpha\beta}(\beta_1-\kappa)\right) - \lambda_1 &= 0 \\
\frac{\partial}{\partial \beta_1}\mathcal{L} &= 0 \\
\iff (1-\delta)r_\beta - 2c_{1,\beta\beta}(\beta_1-\kappa) + 2c_{1,\alpha\beta}(\alpha_1-\alpha_0) - \lambda_2 - \lambda_3\frac{\partial U_D}{\partial \beta_1} &= 0 \\
\iff (1-\lambda_3)\left((1-\delta)r_\beta - 2c_{1,\beta\beta}(\beta_1-\kappa) + 2c_{1,\alpha\beta}(\alpha_1-\alpha_0)\right) - \lambda_2 &= 0 \\
\frac{\partial}{\partial \lambda_1}\mathcal{L} &= 0 \\
\iff -\alpha_1 + \alpha_0 + s_1^2 &= 0 \\
\frac{\partial}{\partial \lambda_2}\mathcal{L} &= 0 \\
\iff -\beta_1 + \kappa + s_2^2 &= 0 \\
\frac{\partial}{\partial \lambda_3}\mathcal{L} &= 0 \\
\iff -U_D + s_2^2 &= 0 \\
\iff -(1-\delta)r^T\gamma_1 + (\gamma_1-\gamma_0)^T C_1 (\gamma_1-\gamma_0) + s_2^2 &= 0
\end{aligned}
$$

Using complementary slackness, we have eight possible options:

1. $s_1 = 0, \lambda_1 > 0, s_2 = 0, \lambda_2 > 0, s_3 \neq 0, \lambda_3 = 0 \to \beta_1^* = \kappa, \alpha_1^* = \alpha_0$.

2. $s_1 = 0, \lambda_1 > 0, s_2 = 0, \lambda_2 > 0, s_3 = 0, \lambda_3 > 0 \to \beta_1^* = \kappa, \alpha_1^* = \alpha_0$. This offers the same candidate as (1).

3. $s_1 = 0, \lambda_1 > 0, s_2 \neq 0, \lambda_2 = 0, s_3 \neq 0, \lambda_3 = 0 \to \alpha_1^* = \alpha_0$, solve equations (1) and (2) for β_1^* and λ_1. Omitting the algebra, this yields:

$$\gamma_1^* = \begin{bmatrix} \alpha_0 \\ \beta_0 + \frac{(1-\delta)r_\beta}{2c_{1\beta\beta}} \end{bmatrix}$$

4. $s_1 = 0, \lambda_1 > 0, s_2 \neq 0, \lambda_2 = 0, s_3 = 0, \lambda_3 > 0 \to \alpha_1^* = \alpha_0$, solve equations (1) and (2) for β_1^* and λ_1. This solution, if it is distinct from the previous solution (3), will always be dominated because it is characterized by 0 utility for G.

5. $s_1 \neq 0, \lambda_1 = 0, s_2 = 0, \lambda_2 > 0, s_3 = 0, \lambda_3 > 0 \to \beta_1^* = \kappa \to$ solve equations (1) and (2) for λ_1 and α_1^*. Omitting algebra, this yields:

$$\gamma_1^* = \begin{bmatrix} \alpha_0 + \frac{(1-\delta)r_\alpha}{2c_{1\alpha\alpha}} - \frac{c_{1\alpha\beta}}{c_{1\alpha\alpha}}\max(0, \theta_D - \beta_0) \\ \max(\beta_0, \theta_D) \end{bmatrix}$$

6. $s_1 \neq 0, \lambda_1 = 0, s_2 = 0, \lambda_2 > 0, s_3 \neq 0, \lambda_3 = 0 \rightarrow$ this solution, if it is distinct from the previous one (5), will always be dominated because it is characterized by 0 utility for G.

7. $s_1 \neq 0, \lambda_1 = 0, s_2 \neq 0, \lambda_2 = 0, s_3 \neq 0, \lambda_3 = 0 \rightarrow \alpha_1^* = \alpha_0$, solve equations (1) and (2) for α_1^*, β_1^*. This is the unconstrained solution. Omitting algebra, this yields: $\gamma_1^* = \gamma_0 + \frac{(1-\delta)}{2} C_1^{-1} r$.

8. $s_1 \neq 0, \lambda_1 = 0, s_2 \neq 0, \lambda_2 = 0, s_3 = 0, \lambda_3 > 0 \rightarrow \beta_1^* = \kappa \rightarrow$ solve equations (1) and (2) for α_1^*, β_1^*. This solution, if it is distinct from (7), will always be dominated by (7) because it is characterized by 0 utility for G.

Thus we have established our four candidates in the proposition statement. To handle the `abstain` scenario, we check each candidate produced in the process above by plugging the strategy to our formula for U_D. If none yield positive utility, then the domain specialist prefers to `abstain`. □

The generalist's subgame perfect equilibrium strategy under regulation. Here we prove Proposition 4.5.

Proof. G's best-response strategy is the value γ_0^* that maximizes G's utility.

$$\gamma_0^*(\delta, \theta_G, \theta_D) \quad = \quad \arg\max_{\gamma_0} U_G(\gamma_0; \delta, \theta_G, \theta_D) \text{ s.t. } U_G \geq 0, U_D \geq 0, \alpha_0 \geq 0, \beta_0 \geq \theta_G.$$

To solve the optimization, we specify the Lagrangian as follows for some multipliers $\lambda \in \mathbb{R}^4$ and a slack variables $s \in \mathbb{R}^4$.

$$\mathcal{L} := \quad \delta r^T \gamma_1 - \gamma_0^T C_0 \gamma_0 - \lambda_1(\alpha_0 - s_1^2) - \lambda_2(\beta_1 - \theta_G - s_2^2) - \lambda_3(U_D - s_3^2) - \lambda_4(U_G - s_4^2).$$

We partially differentiate with respect to each decision variable and each multiplier.

$$\begin{aligned}
\frac{\partial}{\partial \alpha_1}\mathcal{L} &= 0 \\
\iff \delta r_\alpha - 2c_{0,\alpha\alpha}\alpha_0 + 2c_{0,\alpha\beta}\beta_0 - \lambda_1 - \lambda_3\frac{\partial U_G}{\partial \alpha_0} - \lambda_4\frac{\partial U_D}{\partial \alpha_0} &= 0 \\
\iff (1-\lambda_3)\left(\delta r_\alpha - 2c_{0,\alpha\alpha}\alpha_0 + 2c_{0,\alpha\beta}\beta_0\right) - \lambda_1 & \\
-\lambda_4\left((1-\delta)r_\alpha - 2c_{1,\alpha\alpha}(\alpha_1-\alpha_0) + 2c_{1,\alpha\beta}(\beta_1 - \max(\beta_0,\theta_D))\right) &= 0, \\
\frac{\partial}{\partial \beta_1}\mathcal{L} &= 0 \\
\iff \delta r_\beta - 2c_{0,\beta\beta}\beta_0 + 2c_{0,\alpha\beta}\alpha_0 - \lambda_2 - \lambda_3\frac{\partial U_G}{\partial \beta_0} - \lambda_4\frac{\partial U_D}{\partial \beta_0} &= 0 \\
\iff (1-\lambda_3)\left(\delta r_\beta - 2c_{0,\beta\beta}\beta_0 + 2c_{0,\alpha\beta}\alpha_0\right) - \lambda_2 & \\
-\lambda_4\left((1-\delta)r_\beta - 2c_{1,\beta\beta}(\beta_1 - \max(\beta_0,\theta_D) + 2c_{1,\alpha\beta}(\alpha_1-\alpha_0)\right) &= 0, \\
\frac{\partial}{\partial \lambda_1}\mathcal{L} &= 0 \\
\iff -\alpha_0 + s_1^2 &= 0, \\
\frac{\partial}{\partial \lambda_2}\mathcal{L} &= 0 \\
\iff -\beta_0 + s_2^2 &= 0, \\
\frac{\partial}{\partial \lambda_3}\mathcal{L} &= 0 \\
\iff -U_G + s_2^2 &= 0 \\
\iff -\delta r^T\gamma_1 + \gamma_0^T C_1\gamma_0 + s_2^2 &= 0, \\
\frac{\partial}{\partial \lambda_3}\mathcal{L} &= 0 \\
\iff -U_D + s_2^2 &= 0 \\
\iff -(1-\delta)r^T\gamma_1 + (\gamma_1-\gamma_0)^T C_1(\gamma_1-\gamma_0) + s_2^2 &= 0.
\end{aligned}$$

Using complementary slackness, we have sixteen possible options. For brevity, we refer to these options by the constraints they satisfy, where **bold** corresponds to the constraints being activated. The algebra is omitted for exposition; only the candidates yielded are noted for each constraint setting.

1. $\boldsymbol{\alpha_0, \beta_0, U_G, U_D} \rightarrow [0, \theta_G]$.

2. $\boldsymbol{\alpha_0, \beta_0, U_G}, U_D \rightarrow [0, \theta_G]$

3. $\boldsymbol{\alpha_0, \beta_0}, U_G, U_D \rightarrow [0, \theta_G]$

4. $\boldsymbol{\alpha_0, \beta_0}, U_G, \boldsymbol{U_D} \rightarrow [0, \theta_G]$

5. $\boldsymbol{\alpha_0}, \beta_0, \boldsymbol{U_G, U_D} \rightarrow \begin{bmatrix} 0 \\ \frac{\delta r_\beta}{2c_{0\beta\beta}} \end{bmatrix}$

6. $\boldsymbol{\alpha_0}, \beta_0, \boldsymbol{U_G}, U_D \rightarrow \begin{bmatrix} 0 \\ \frac{\delta r_\beta}{2c_{0\beta\beta}} \end{bmatrix}$
7. $\boldsymbol{\alpha_0}, \beta_0, U_G, U_D \rightarrow \begin{bmatrix} 0 \\ \frac{\delta r_\beta}{2c_{0\beta\beta}} \end{bmatrix}$
8. $\boldsymbol{\alpha_0}, \beta_0, U_G, \boldsymbol{U_D} \rightarrow$ One of three along $U_D = 0$ curve.
9. $\alpha_0, \boldsymbol{\beta_0}, \boldsymbol{U_G}, \boldsymbol{U_D} \rightarrow \gamma_0^* = \begin{bmatrix} \frac{\delta r_\alpha}{2c_{0\alpha\alpha}} - \frac{c_{0\alpha\beta}}{c_{0\alpha\alpha}}\theta_G \\ \theta_G \end{bmatrix}.$
10. $\alpha_0, \boldsymbol{\beta_0}, \boldsymbol{U_G}, U_D \rightarrow \begin{bmatrix} \frac{\delta r_\alpha}{2c_{0\alpha\alpha}} - \frac{c_{0\alpha\beta}}{c_{0\alpha\alpha}}\theta_G \\ \theta_G \end{bmatrix}$
11. $\alpha_0, \boldsymbol{\beta_0}, U_G, U_D \rightarrow \begin{bmatrix} \frac{\delta r_\alpha}{2c_{0\alpha\alpha}} - \frac{c_{0\alpha\beta}}{c_{0\alpha\alpha}}\theta_G \\ \theta_G \end{bmatrix}$
12. $\alpha_0, \boldsymbol{\beta_0}, U_G, \boldsymbol{U_D} \rightarrow$ Two of three along the $U_D = 0$ curve.
13. $\alpha_0, \beta_0, \boldsymbol{U_G}, \boldsymbol{U_D} \rightarrow \frac{\delta}{2} C_0^{-1} r$.
14. $\alpha_0, \beta_0, \boldsymbol{U_G}, U_D \rightarrow \frac{\delta}{2} C_0^{-1} r$
15. $\alpha_0, \beta_0, U_G, U_D \rightarrow \frac{\delta}{2} C_0^{-1} r$
16. $\alpha_0, \beta_0, U_G, \boldsymbol{U_D} \rightarrow$ Three of three along the $U_D = 0$ curve.

Thus we have established our four candidates in the proposition statement. To handle the `abstain` scenario, we check each candidate produced in the process above by plugging the strategy to our formula for U_G. If none yield positive utility, then the generalist prefers to `abstain`. □

9 Helper Lemmas and Analysis

Here we write out helper Lemmas and analysis for our proofs concerning backfiring and mutualism.

Lemma 9.1. *In the AI regulation game with quadratic costs, any player's utility is strictly concave if and only if their cost matrix is positive definite.*

Proof. The generalist utility function is given by $U_G = \delta r^T \gamma_1 - \gamma_0^T C_0 \gamma_0$. Observe this is twice differentiable everywhere. Thus the function is strictly concave in α_0, β_0 if and only if its Hessian derivative is negative definite. We compute the Hessian as follows:

$$H := \begin{bmatrix} \frac{\partial^2 U_G}{\partial \alpha_0^2} & \frac{\partial^2 U_G}{\partial \alpha_0 \partial \beta_0} \\ \frac{\partial^2 U_G}{\partial \beta_0 \partial \alpha_0} & \frac{\partial^2 U_G}{\partial \beta_0^2} \end{bmatrix} = -2C_0.$$

This matrix is negative definite if and only if C_0 is positive definite.

The proof for the domain specialist follows the same steps. □

Lemma 9.2. *In any AI regulation game with separable quadratic costs, if there is no regulation, both players will invest a non-zero amount in each attribute.*

Proof. By Lemma 9.1, we are given that the utilities are strictly concave. Thus, the proof consists of showing that 1) the utility function is greater than or equal to 0 at the origin point of zero investment and 2) the gradient points towards the interior of the feasible set everywhere along the boundaries.

Here we prove the two conditions for U_G:

1. $U_G(\alpha_0 = 0, \beta_0 = 0) = \delta r^T \vec{0} - 0 = 0$
2. We prove this for each constraint, $\alpha_0 \geq 0, \beta_0 \geq 0$:
    - $\frac{\partial U_G}{\partial \alpha_0}\big|_{\beta_0=0} = \delta r_\alpha - 2c_{0,\alpha\alpha}\alpha_0 = \delta r_\alpha - 0 > 0.$
    - $\frac{\partial U_G}{\partial \beta_0}\big|_{\alpha_0=0} = \delta r_\beta - 2c_{0,\beta\beta}\beta_0 = \delta r_\beta - 0 > 0.$

Here we prove the two conditions for U_D:

1. $U_D(\alpha_i = 0, \beta_i = 0) = (1 - \delta) r^T \gamma_0 - 0 \geq 0$
2. We prove this for each constraint, $\alpha_i \geq \alpha_0, \beta_i \geq \beta_0$:
    - $\frac{\partial U_D}{\partial \alpha_i}\big|_{\beta_i=\beta_0} = (1 - \delta) r_\alpha - 2c_{i,\alpha\alpha}(\alpha_1 - \alpha_0) = (1 - \delta) r_\alpha > 0.$
    - $\frac{\partial U_G}{\partial \beta_0}\big|_{\alpha_i=\alpha_0} = (1 - \delta) r_\beta - 2c_{i,\beta\beta}(\beta_i - \beta_0) = \delta r_\beta - 0 > 0.$

□

10 Proving the Backfiring Result

Below we prove the Theorem 6.1.

Proof. Assume $\theta_G = 0$ for the entire proof. By Remark 4.3, we're given that the players commit to their unconstrained strategy in equilibrium. These were solved in Propositions 4.4 and 4.5. Thus we have the following player's strategies under no regulation for this setting:

$$\gamma_0^A = \frac{\delta}{2} C_0^{-1} r, \quad \gamma_1^A = \frac{1 - \delta}{2} C_1^{-1} r. \tag{1}$$

Our strategy is to show that G's unconstrained, no-regulation optimum becomes dominated in the presence of regulation targeting D, which we choose to be arbitrarily close to β_1^A.

Notation. Before we proceed, we introduce some additional notation. Define the set S to be all feasible pairs of strategies (γ_0, γ_1). 'Feasible' here means those strategies which leave both G and D with non-negative utility. We use the subscript S_{θ_D} to track the particular regulatory threshold. The feasible pairs of strategies in the unregulated game is given by S_0, and the feasible pairs of strategies in a game with threshold $\theta_D = 1.5$ is denoted $S_{1.5}$. We may refer to the unregulated game with the superscript A (for anarchy), e.g. β_1^A refers to the unregulated safety level. Observe that any set of tuples S_θ can be separated into two mutually exclusive and collectively exhaustive sets:

- S_θ^{MC} (for minimally compliant) is the set of all tuples where D's best response has safety $\beta_1^* = \theta_D$.
- S_θ^{C} (for contribute) is the set of all tuples where D's best response has safety $\beta_1^* > \theta_D$.

Now, we provide a sequence of lemmas, with the purpose of establishing the intuition that *all we must do is find some $\epsilon > 0$ and some strategy $\beta_0^R \neq \beta_0^A$ such that G prefers β_0^R to β_0^A and D minimally complies.*

Lemma 10.1. *For any threshold $\theta_D > 0$, $S_\theta^C \subset S_0$.*

Proof. $S_0 = S_0^{\mathrm{MC}} \cup S_0^{\mathrm{C}} = S_0^{\mathrm{MC}} \cup \left(\bigcup_{t=0}^{\infty} S_t^{\mathrm{C}}\right) \supset S_\theta^{\mathrm{C}}$. □

Lemma 10.2. *If $\theta_D \geq \beta_1^A$, backfiring is impossible.*

Proof. Assume for contradiction that $\theta_D^* \geq \beta_1^A$ and backfiring occurs. Backfiring would imply $\beta_1(\theta_D = \theta_D^*) < \beta_1(\theta_D = 0) = \beta_1^A$. However, this would violate the regulation, which we're given is greater than β_1^A. Hence we've already established the contradiction. □

Lemma 10.3. *Given a threshold θ, backfiring can occur only if the strategies $(\gamma_0, \gamma_1) \in S_{\theta_D}^{MC}$.*

Proof. We have established $S_{\theta_D} = S_{\theta_D}^{\mathrm{MC}} \cup S_{\theta_D}^{\mathrm{C}}$, so the proof will show that the strategies in S_θ^{C} can never exhibit backfiring. This would imply, if backfiring occurs over the feasible set of strategies S_θ, it is only possible for strategies in S_θ^{MC}. The proof proceeds, first for all values $\theta_D \geq \beta_1^A$, and then for all values $\theta < \beta_1^A$.

- For $\theta_D \geq \beta_1^A$, backfiring is impossible generally, as established in Lemma 10.2.
- For $\theta_D < \beta_1^A$, start by observing that the anarchy solution (γ_0^A, γ_1^A) is always feasible. This solution is the strategy tuple that maximizes the utility of G over S_0. Lemma 10.1 tells us that this set, S_0, contains all sets of regulated strategies where the players contribute: $S_{\theta_D}^{\text{contribute}} \subset S_0$. Thus: $(\gamma_0^A, \gamma_1^A) := \sup_{U_G} S_0 \succeq_G S_0 \supset S_{\theta_D}^{\text{contribute}} \rightarrow (\gamma_0^A, \gamma_1^A) \succeq_G S_{\theta_D}^{\mathrm{C}}$. Thus the anarchy solution is feasible and dominates all strategies in $S_{\theta_D}^{\text{contribute}}$.

This completes the proof, and demonstrates that if backfiring is ever to occur, it will exhibit strategies that are *minimally compliant* with the regulation. □

Backfiring is a regulation yielding lower safety than β_0^A. The claims above state that backfiring cannot occur if $\theta_D > \beta_1^A$ and can only occur if the domain specialist minimally complies. As an immediate corollary, we can claim that backfiring occurs *if and only if* there is a regulation $\theta_D < \beta_1^A$ such that the strategies $(\gamma_0(\theta_D), \gamma_1(\theta_D)) \in S_{\theta_D}^{MC}$.

Lemma 10.4. *For a given regulation $\theta_D < \beta_1^A$ in our setting, if G prefers any minimally compliant strategy γ_0' to γ_0^A, then G's optimal strategy $\gamma_0^* \in S_{\theta_D}^{MC}$ and the regulation backfires.*

Proof. We're given γ_0^A is optimal over S_0. By Lemma 10.1, $S_{\theta_D}^C \subset S_0$. Since $\theta_D < \beta_1^A$, γ_0^A remains feasible. The only new strategies available to G are those in $S_{\theta_D}^{MC}$. Thus, if we denote utility-domination using $\succ$, we have $\gamma_0' \succ \gamma_0^A \succeq g \forall g \in S_{\theta_D}^C$. This implies G's optimal strategy γ_0^* is either γ_0' or otherwise belongs to $S_{\theta_D}^{MC}$. □

Thus our task is to find some regulation θ_D and some strategy γ_0' such that $U_G(\gamma_0') > U_G(\gamma_0^A)$.

Lemma 10.5. *For small $\epsilon > 0$, if the given conditions are met, the following G strategy dominates no regulation:*

$$\gamma_0' = \begin{bmatrix} \frac{\delta r_\alpha}{2c_{0,\alpha\alpha}} \left(\beta_0^A - 2\epsilon\right) \\ \beta_0^A - 2\epsilon \end{bmatrix}$$

Proof. Equation 1 give us G and D's strategies under no regulation. Given G's candidate strategy stated in the Lemma, we compute D's best response. Observe this must be a minimally-compliant best response, because G's strategy was constructed to be a difference $\beta_0^A + \epsilon$ from D's regulatory floor. Thus, by Proposition 4.4, we have:

$$\gamma_1' = \begin{bmatrix} \alpha_0' + \frac{(1-\delta)}{2c_{1\alpha\alpha}} - \frac{c_{1,\alpha\beta}}{c_{1\alpha\alpha}}\theta_D \\ \theta_D \end{bmatrix}$$

We compare G's utility in the two scenarios:

1. $(\gamma_0', \gamma_1') \to U_G' = \delta\left(r_\alpha \alpha_1' + r_\beta \beta_1'\right) - c_{0,\alpha\alpha}(\alpha_0')^2 - 2c_{0,\alpha\beta}\alpha_0'\beta_0' - c_{0,\beta\beta}(\beta_0')^2$
2. $(\gamma_1^A, \gamma_1^A) \to U_G^A = \delta\left(r_\alpha \alpha_1^A + r_\beta \beta_1^A\right) - c_{0,\alpha\alpha}(\alpha_0^A)^2 - 2c_{0,\alpha\beta}\alpha_0^A\beta_0^A - c_{0,\beta\beta}(\beta_0^A)^2$

We compute the difference $\Delta U_G = U_G' - U_G^A$. We expand both terms and take the limit as $\epsilon \searrow 0$ to get the following:

$$\lim_{\epsilon \searrow 0} \Delta U_G = \frac{\delta(1-\delta)r_\beta}{2c_{1\beta\beta}} \left(r_\beta - \frac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}} r_\alpha\right)$$

A sufficient condition for this quantity being positive is stated below. The reason is all terms outside the parentheses are given as positive.

$$r_\beta > \frac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}} r_\alpha.$$

Notice the above condition is given as it is one of the conditions in remark 4.3.[7] □

Thus, we have shown that for small positive ϵ, the generalist prefers the backfiring strategy to the unconstrained optimum γ_0^A. By Lemma 10.4, the optimal regulated strategy is an element in $S_{\theta_D}^{MC}$ and the regulation backfires. □

11 Proof of the mutualism result

Here we prove Theorem 6.2, that for a swath of games there exists a set of regluations that mutually improve the player's utilities.

[7] This is also the condition for having a non-zero safety investment when costs are convex, and intuitively, backfiring is impossible when safety investment is zero.

Proof. Observe that we only have to provide a single instance of regulation that does better than the unregulated optimal γ_0^A, γ_1^A to show that there exists a Pareto improvement effect of regulation. We consider the following minimal-compliance strategies (using Proposition 4.4 and 4.5):

$$\gamma_0' = \begin{bmatrix} \frac{\delta r_\alpha}{2c_{0,\alpha\alpha}} - \frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\theta_G \\ \theta_G \end{bmatrix}, \gamma_1' = \begin{bmatrix} \alpha_0 + \frac{(1-\delta)r_\alpha}{2c_{1,\alpha\alpha}} - \frac{c_{1,\alpha\beta}}{c_1\alpha\alpha}\theta_D \\ \theta_D \end{bmatrix}.$$

Observe these are feasible because they are compliant and, for small ϵ, the performance investment is positive.

Lemma 11.1. *For the specified conditions,* $U_G(\gamma_0', \gamma_1') > U_G(\gamma_0^A, \gamma_1^A)$

Proof. Start by computing the change in the generalist's performance and safety investments between these strategies. The change in safety investment is simply $\Delta\beta_0 = \theta_G - \beta_0^A = \epsilon$. The change in performance investment is given by:

$$\begin{aligned}
\Delta\alpha_0 = \quad & \frac{\delta r_\alpha}{2c_{0,\alpha\alpha}} - \frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\left(\frac{\delta}{2\det C_0}(-c_{0,\alpha\beta}r_\alpha + c_{0,\alpha\alpha}r_\beta + \epsilon)\right) - \frac{\delta}{2\det C_0}\left(c_{0\beta\beta}r_\alpha - c_{0\alpha\beta}r_\beta\right) \\
= \quad & \frac{\delta r_\alpha}{2c_{0,\alpha\alpha}} + \frac{c_{0,\alpha\beta}^2}{c_{0,\alpha\alpha}}\frac{\delta r_\alpha}{2\det C_0} - \cancel{c_{0,\alpha\beta}\frac{\delta}{2\det C_0}r_\beta} + \frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\epsilon - \frac{\delta}{2\det C_0}c_{0,\beta\beta}r_\alpha + \cancel{c_{0,\alpha\beta}\frac{\delta}{2\det C_0}r_\beta} \\
= \quad & \left(\frac{\delta}{2c_{0,\alpha\alpha}} + \frac{c_{0,\alpha\beta}^2\delta}{c_{0,\alpha\alpha}2\det C_0} - \frac{\delta c_{0,\beta\beta}}{2\det C_0}\right)r_\alpha + \frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\epsilon \\
= \quad & \frac{\delta r_\alpha}{2}\left(\frac{1}{c_{0,\alpha\alpha}} + \frac{c_{0,\alpha\beta}^2}{c_{0,\alpha\alpha}(c_{0,\alpha\alpha}c_{0,\beta\beta} - c_{0,\alpha\beta}^2)} - \frac{c_{0,\beta\beta}}{c_{0,\alpha\alpha}}\right) + \frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\epsilon \\
= \quad & \frac{\delta r_\alpha}{2}\cancel{\left(\frac{\det C_0 + c_{\alpha\beta}^2 - c_{0,\alpha\alpha}c_{0,\beta\beta}}{c_{0,\alpha\alpha}\det C_0}\right)} + \frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\epsilon \\
= \quad & \frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\epsilon.
\end{aligned}$$

By the same logic, we solve for the change in the players' strategies. First, $\Delta\beta_1 = \beta_1' - \beta_1^A = \beta_1^A + 2\epsilon - \beta_1^A = 2\epsilon$. The change in performance is given by:

$$\begin{aligned}
\Delta\alpha_1 = \quad & \alpha_1' - \alpha_1^A \\
= \quad & \left[\alpha_0' + \frac{(1-\delta)r_\alpha}{2c_{1,\alpha\alpha}} - \frac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}}\left(\theta_D - \beta_0'\right)\right] - \left[\alpha_0^A + \frac{(1-\delta)}{2\det C_1}(c_{1,\beta\beta}r_\alpha - c_{1,\alpha\beta}r_\beta)\right] \\
= \quad & \Delta\alpha_0 + \frac{(1-\delta)r_\alpha}{2c_{1,\alpha\alpha}} - \frac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}}\left(\frac{(1-\delta)}{2\det C_1}\left(-c_{1,\alpha\beta}r_\alpha + c_{1,\alpha\alpha}r_\beta\right) + \epsilon\right) - \frac{(1-\delta)}{2\det C_1}\left(c_{1,\beta\beta}r_\alpha - c_{1,\alpha\beta}r_\beta\right) \\
= \quad & \frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\epsilon + \frac{(1-\delta)r_\alpha}{2c_{1,\alpha\alpha}} - \frac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}}\left(\frac{(1-\delta)}{2\det C_1}\left(-c_{1,\alpha\beta}r_\alpha + c_{1,\alpha\alpha}r_\beta\right) + \epsilon\right) - \frac{(1-\delta)}{2\det C_1}\left(c_{1,\beta\beta}r_\alpha - c_{1,\alpha\beta}r_\beta\right) \\
= \quad & \epsilon\left(\frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}} - \frac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}}\right).
\end{aligned}$$

The change in G's cost is given by:

$$\begin{aligned}
\Delta(\text{G's cost}) = \quad & \begin{bmatrix} \Delta\alpha_0 \\ \Delta\beta_0 \end{bmatrix}^T C_0 \begin{bmatrix} \Delta\alpha_0 \\ \Delta\beta_0 \end{bmatrix} \\
= \quad & c_{0,\alpha\alpha}\left(\frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\right)^2\epsilon^2 + 2\left(\frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}}\epsilon\right)\epsilon + c_{0\beta\beta}\epsilon^2
\end{aligned}$$

Notice these are all ϵ^2 terms, meaning as ϵ is brought to very small positive values, they approach zero at an exponential rate. The contribution to G's revenue is given by:

$$\begin{aligned}\Delta(\text{G's revenue}) = & \quad \delta(r_\alpha \Delta\alpha_1 + r_\beta \Delta\beta_1) \\ = & \quad \delta\left(r_\alpha \epsilon \left(\tfrac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}} - \tfrac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}}\right) + r_\beta 2\epsilon\right)\end{aligned}$$

Notice these are terms of ϵ, whereas the cost effects are solely terms of ϵ^2. Therefore, for sufficiently small ϵ, we say:

$$\lim_{\epsilon \searrow 0} \Delta U_G = \delta\left(r_\alpha \epsilon \left(\frac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}} - \frac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}} + 2r_\beta \epsilon\right)\right)$$

Using the given conditions, we know:

$$\begin{aligned}\lim_{\epsilon \searrow 0} \Delta U_G > 0 \iff & \quad \delta\left(r_\alpha \epsilon \left(\tfrac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}} - \tfrac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}}\right) + 2r_\beta \epsilon\right) > 0 \\ \iff & \quad r_\alpha \left(\tfrac{c_{0,\alpha\beta}}{c_{0,\alpha\alpha}} - \tfrac{c_{1,\alpha\beta}}{c_{1,\alpha\alpha}}\right) + 2r_\beta > 0 \\ \iff & \quad \tfrac{r_\alpha c_{0,\alpha\beta}}{r_\beta c_{0,\alpha\alpha}} - \tfrac{r_\alpha c_{1,\alpha\beta}}{r_\beta c_{1,\alpha\alpha}} > -2.\end{aligned}$$

Our conditions strictly bound the absolute value of both terms on the left hand side below 1, so this completes the Lemma's proof. □

Lemma 11.2. *For the specified conditions,* $U_D(\gamma_0', \gamma_1') > U_D(\gamma_0^A, \gamma_1^A)$

The limiting effect on D's revenue is calculated exactly the same way as above, except that the revenue expression is multiplied by $(1-\delta)$ instead of δ.

This completes the proof, as both players are better off under the regulation. □