Fine-Tuning Games: Bargaining and Adaptation for General-Purpose Models

Benjamin Laufer
Cornell Tech
New York, New York, USA
bdl56@cornell.edu

Jon Kleinberg*
Cornell University
Ithaca, New York, USA
kleinberg@cornell.edu

Hoda Heidari*
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
hheidari@andrew.cmu.edu

ABSTRACT

Recent advances in Machine Learning (ML) and Artificial Intelligence (AI) follow a familiar structure: A firm releases a large, pretrained model. It is designed to be adapted and tweaked by other entities to perform particular, domain-specific functions. The model is heralded as 'general-purpose,' meaning it can be transferred to a wide range of downstream tasks, in a process known as *adaptation* or *fine-tuning*. Understanding this process – the strategies, incentives, and interactions involved in the development of AI tools – is crucial for making conclusions about societal implications and regulatory responses, and may provide insights beyond AI about general-purpose technologies. We propose a model of this adaptation process. A Generalist brings the technology to a certain level of performance, and one or more Domain specialist(s) adapt it for use in particular domain(s). Players incur costs when they invest in the technology, so they need to reach a bargaining agreement on how to share the resulting revenue before making their investment decisions. We find that for a broad class of cost and revenue functions, there exists a set of Pareto-optimal profit-sharing arrangements where the players jointly contribute to the technology. Our analysis, which utilizes methods based on bargaining solutions and sub-game perfect equilibria, provides insights into the strategic behaviors of firms in these types of interactions. For example, profit-sharing can arise even when one firm faces significantly higher costs than another. We show that any potential domain specialization will either *contribute*, *free-ride*, or *abstain* in their uptake of the technology, and provide conditions yielding these different responses.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Adaptation; bargaining; general-purpose technology; fine-tuning

ACM Reference Format:
Benjamin Laufer, Jon Kleinberg, and Hoda Heidari. 2024. Fine-Tuning Games: Bargaining and Adaptation for General-Purpose Models. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3589334.3645366

*Equal contribution

WWW '24, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0171-9/24/05.
https://doi.org/10.1145/3589334.3645366

1 INTRODUCTION

Large-scale AI models have garnered a great deal of excitement because they are considered to be *general purpose* [11, 18, 22, 40, 41, 55]. Some have referred to these technologies as *foundation models* [4, 5, 19] because they are designed as massive, centralized models that support potentially many downstream uses. For example, Bommasani et al. [4] write, "a foundation model is itself incomplete but serves as the common basis from which many task-specific models are built via adaptation." There is palpable excitement about these technologies. But to turn their potential into actual use and impact, one needs to specialize, tweak, and evaluate the technology for particular application domains. This process takes various names, including *adaptation* [43] and, in some contexts, *fine-tuning* [33, 50, 59].

Notably, the process of adapting a technology involves multiple parties. Technology teams developing ML and AI technologies rely on outside entities to adapt, tweak, transfer, and integrate the general-purpose model. This dynamic suggests a latent strategic interaction between producers of a foundational, general-purpose technology and specialists considering whether and how to adopt the technology in a particular context. Understanding this interaction is necessary to study the social, economic, and regulatory consequences of introducing the technology.

This paper employs methods from economic theory to model and analyze this interaction. We put forward a model of fine-tuning where the interaction between two agents, a generalist and a domain-specialist, determines how they'll bring a general-purpose technology to market (Figure 1). The result of this interaction is a domain-adapted product that offers a certain level of *performance* to consumers, in exchange for a certain level of surplus revenue for the producers. Crucially, the producers must decide how to distribute the surplus, and engage in a bargaining process in advance of making their investment decisions. An immediate intuition might be to divide the surplus based on contribution to the technology — however, this is one of many potential bargaining solutions, each with different normative assumptions and implications for the technology's performance and the distribution of utility.

Through this analysis, we uncover several general principles that apply not just to today's AI technologies, but to a potentially wide swath of models that exhibit a similar structure — i.e., developed for general use and adapted to one or more domains to produce revenue. Thus, even as these technologies improve and develop, our proposed model of fine-tuning may continue to describe how they may be adapted for real-world use(s). Further, some of our findings apply to other general-purpose technologies outside the AI context. For example, cloud computing infrastructure enables a number of consumer-facing services that use web hosting, database services,

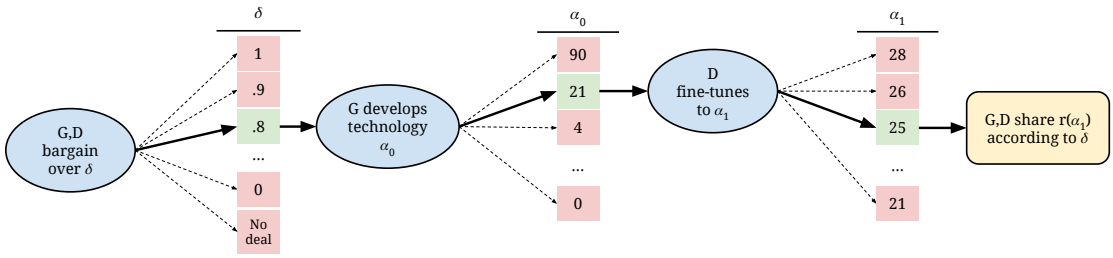

Figure 1: An illustration of the fine-tuning game. In the first step, players bargain over the revenue-sharing agreement δ. In this example, they agree that G will receive 80% of the revenue and D will receive 20%. In the second step, G develops the technology to performance level $\alpha_0 = 21$. In the third step, D 'fine-tunes' the technology to $\alpha_1 = 25$. If the players collectively receive revenue of 25, they would share so that G receives 20 and D receives 5.

and other on-demand computing resources. Additive manufacturing (e.g., 3D printing) requires the production of a general-purpose technology that other entities use to create valuable products in particular domains. Digital marketplaces, too, are general market-making technologies that enable specialists (vendors) to sell goods, subject to an agreement over surplus.

Our main conceptual contribution is modeling the adaptation process as a **multi-stage game** consisting of (1) a **bargaining process** between a general-purpose technology producer (G) and one or more domain specialists (D), and (2) two additional stages for G and D to invest in performance, respectively (see Figure 1). Both players bargain over how to share revenue, and each takes a turn contributing to the technology's performance before it reaches the market. Within the set of Pareto-optimal revenue-sharing agreements, we introduce a number of *bargaining solutions* that represent potential arrangements for how entities involved in AI's development should distribute profit and effort. These bargaining solutions can be thought of as diverse normative proposals for how to appropriately distribute welfare.

Our analysis consists in deriving the sub-game perfect equilibrium strategies, identifying the set of Pareto-optimal bargaining agreements, and then solving for various bargaining solutions. Even in the presence of significant cost differentials, we find bargaining leads to profit-sharing agreements because specialists can leverage their power to exit the deal, reducing the reach of the technology — or, in the case of one specialist, preventing the technology from being produced altogether. For fine-tuning games with a somewhat general set of cost and revenue functions, we develop a method for identifying Pareto-optimal bargains. A significant, high-level take-away from our analysis is a characterization of the specialist fine-tuning strategy. We find that any potential adaptor of a technology falls into one of three groups: **contributors**, who invest effort before selling the technology; **free-riders**, who sell the technology without investing any additional effort; and **abstainers**, who do not enter any fine-tuning agreement and opt not to bring the technology to their particular domain. It turns out, using only marginal information about a domain (0th- and 1st-order approximations of cost and revenue), it is possible to reliably determine which strategy the adaptor will take for a notably broad set of scenarios and cost and revenue functions (Section 4.1).

Some have suggested that scholarship on AI and data-driven technologies focuses predominantly on the technical developments without situating these developments in political economy (though notable exceptions exist) [1, 7, 9, 52, 56]. We propose a model that accounts for the different interests and interactions involved in the development of new, general-purpose AI technology. Our model enables analysis on how these interactions affect market outcomes like performance in practice. Understanding these interactions may also inform future regulation of harms that can arise from large-scale ML technologies.

1.1 Related Work

There exists a considerable body of literature on methods for fine-tuning and adapting general ML models. Our work leverages economic theory to understand the incentives and strategies that determine how these general-purpose technologies develop.

Approaches to fine-tuning. New applications of ML often involve leveraging an existing model to a specific task, in a process known as transfer learning [60]. As a result, a variety of broad and flexible base models have been developed ('pretrained') for downstream adaptation to particular tasks. These include large language models [8, 12, 28] and visual models [44, 58]. Fine-tuning is an approach where new data and training methods are applied to a pretrained base model to improve performance on a domain-specific task [13]. Fine-tuning often consists of several steps: (1) gathering, processing and labeling domain-specific data, (2) choosing and adjusting the base model's architecture (including number of layers [54] and parameters [48]) and the appropriate objective function [23], (3) Updating the model parameters using techniques like gradient descent, and (4) evaluating the resulting model and refining if necessary.

Economic models of general-purpose technology production. Several lines of work in growth economics address the development and diffusion of general-purpose technologies (or GPTs). Bresnahan [6] provides a general survey of this concept. Jovanovic and Rousseau [30] offers a historic account of technologies such as electricity and information technology as GPTs with major impacts on the United States economy. Scholars have examined the effects of factors such as knowledge accumulation, entrepreneurial activity, network effects, and sectoral interactions on the creation of GPTs [27]. The model presented here abstracts away the forces giving rise to the invention of general-purpose technologies, and instead focuses on the later-stage decision of when (or at what performance level) to release the GPT to market for domain-specialization.

Some have suggested that general-purpose technologies create the need for new business models that describe their impact on individual sectors [36]. Gambardella and McGahan [21] propose one such model of domain adaptation for a general-purpose technology that is based on revenue sharing — however, they do not use bargaining or multi-stage strategy to describe how the technology is developed and brought to market. Our notion of *performance* as it relates to model technologies is inspired by economic models of product innovation [10, 53].

Bargaining and joint production. This work draws from a long line of research on welfare economics and cooperative game theory devoted to understanding how agents reach agreements when their interests are intertwined [15]. Methods have been developed for finding an optimal set of agreements (e.g., contract curves [17] and cores [2, 49]) in exchange economies, where agents can trade goods. When parties must reach an agreement to jointly produce a product, they often engage in a *bargain* – we discuss bargaining further in Section 2.1. Existing empirical work observes how real people or firms bargain, and measures how close these agreements are to those proposed by theorists [20, 51]. A setting with similar models is the development *supply chains* where different firms or entities negotiate over how much effort they each invest and how much profit they each receive [16, 57]. A related body of work is referred to as the *hold-up problem* [46]. This work analyzes settings where two (or more) agents negotiate over an *incomplete* contract and distribute surplus [26]. In these models, after an initial agreement, players are able to re-negotiate and alter parts of the contract, yielding shifts in strategy.

Game Theory and ML. Our paper contributes to a line of work using game-theoretic methods to describe the development of (and responses to) ML models [25, 32, 37, 42] and their societal implications [29, 35, 38]. Donahue and Kleinberg [14] explore a game-theoretic setting where agents may voluntarily take part in a federated learning arrangement. Their setting is a coalitional game among parties that all move simultaneously, whereas ours is a sequential game that involves parties with different roles in the process. Focusing on digital platforms, Hardt et al. [24] describe interactions between a firm implementing an ML algorithm and collectives of users who manipulate their data to influence the algorithm.

2 A MODEL OF FINE-TUNING

In this section, we put forward a model of fine-tuning a data-driven technology for use in a domain-specific context. The technology is developed in two steps: First, a general-purpose producer develops a technology up to a certain level of performance. Then, a domain-specific producer decides whether to adopt the technology, and how much to invest in the technology to further improve its performance beyond the general-purpose baseline. After these steps, the two entities share a payout.

Generalist. Player G (for General-purpose producer) is the first to invest in the technology's performance, and brings the performance level to $\alpha_0 \in \mathbb{R}^+$. G is motivated to invest in the technology because, ultimately, the technology's performance level determines the revenue G earns.

Domain Specialist. After investing in the technology, G can offer the technology to a domain-specialist, denoted D, who fine-tunes the model to their specific use case. If D and G enter an agreement, D will invest in improving the technology's performance from α_0 to $\alpha_1 \in \mathbb{R}^+$ where $\alpha_1 \geq \alpha_0$.

Revenue and costs. The technology's *performance*, α_1, determines the total revenue that can be gained from fine-tuning the technology in that domain. In particular, we assume there is a monotonic function $r : \mathbb{R}^+ \to \mathbb{R}^+$ such that $r(\alpha_1)$ is the total revenue generated by performance level α_1. Unless otherwise specified, we assume $r(\cdot)$ is the identity function, that is, the total revenue brought by the technology is α_1. The cost associated with producing α_1 requires considering the two steps involved with developing the technology: general production and fine-tuning. We say that G faces cost function $\phi_0(\alpha_0) : \mathbb{R}^+ \to \mathbb{R}^+$ to produce a general technology at performance-level α_0. D faces cost function $\phi_1(\alpha_1; \alpha_0) : \mathbb{R}^+ \to \mathbb{R}^+$ to bring the technology from performance α_0 to performance α_1. We assume these cost functions are publicly known.

The fine-tuning game. The players are G and D. In deciding whether to purchase the technology, D negotiates revenue sharing with G. G and D share revenue $r(\alpha_1)$ according to a bargaining parameter $\delta \in [0, 1]$. At the end of the game, G receives $\delta r(\alpha_1)$ in revenue, and D receives $(1 - \delta)r(\alpha_1)$. The model fine-tuning game consists in each player deciding their level of investment and collectively bargaining to decide δ. The game proceeds as follows:

(1) G and D negotiate bargaining coefficient $\delta \in [0, 1]$.
(2) G invests in a general-purpose technology, subject to cost $\phi_0(\alpha_0)$, yielding performance-level α_0.
(3) D fine-tunes the technology, subject to cost $\phi_1(\alpha_1; \alpha_0)$, yielding performance-level α_1.

The steps of the game are illustrated in Figure 1. Players earn the following utilities, defined as revenue share minus cost:

$$U_G(\delta) := \delta r(\alpha_1) - \phi_0(\alpha_0), \;\; U_D(\delta) := (1-\delta) r(\alpha_1) - \phi_1(\alpha_1; \alpha_0). \quad (1)$$

If the players do not agree to a feasible bargain $\delta \in [0, 1]$, then the bargaining outcome is referred to as *disagreement*. In this scenario, the generalist receives d_0 and the specialist receives d_1. We assume, unless otherwise specified, that the disagreement scenario is described by $d_0 = d_1 = 0$.

2.1 Primer on Bargaining Games

Bargaining games are a potentially useful method for computer science research. In this section we include a primer on these methods before demonstrating their use in our model.

A bargain is a process for identifying joint agreements between two or more agents on how to share payoff. The ***Bargaining Problem***, formalized by [39], consists of two players that must jointly decide how to share surplus profit. The problem consists of a set of feasible agreements and a 'disagreement' alternative, which specifies the utilities players receive if they do not come to an agreement.

Bargaining solutions are established ways to select among candidate agreements on how to share surplus. Different bargaining solutions, proposed over the years by mathematicians and economists, aim to satisfy certain desiderata like fairness, Pareto optimality, and utility-maximization. Typically, solving for bargaining solutions consists in defining some measure of *joint utility* between players (e.g. take the sum, product, or minimum of the players'

utilities). The feasible, Pareto-optimal solution that maximizes this joint utility is known as a ***bargaining solution***.

Bargaining solutions are normative: they provide guidelines for how surplus payoffs should be distributed. Solutions are inspired by moral theories like utilitarianism (which aims to maximize the sum of utilities) and egalitarianism (which aims to maximize the worst-off agent). We demonstrate the use of bargaining solutions in the subsequent sections.

2.2 Pareto-Optimal Bargains

Our model of the fine-tuning process unfolds in two stages: the first stage is a bargain where the players must jointly agree on δ, and the second stage is a sequential game where the players make decisions individually in order (i.e., G moves first and D moves second). Our analysis will identify the players' equilibrium strategies and a variety of bargaining solutions δ with different welfare implications. In order to derive solutions, it is important to define *Pareto dominance* and *Pareto efficiency*. Since our analysis relies on these concepts, in this section, we state our first result deriving the set of Pareto-optimal solutions for a general set of cost and revenue functions. We begin by defining relevant concepts.

Definition 2.1 (Pareto-dominant agreements). *A bargaining agreement δ_a **Pareto-dominates** an alternative agreement $\delta_b \neq \delta_a$ iff at least one player gains utility by switching from δ_b to δ_a, and no players lose utility.*

Definition 2.2 (Pareto-optimal agreements). *A **Pareto-optimal** agreement is one where no alternative agreement would improve the utility of one player without decreasing the utility of the other player. In other words, it is an agreement that is not Pareto-dominated by any other agreement.*

Definition 2.3 (Strictly Unimodal Function). *A function $f : \mathbb{R} \to \mathbb{R}$ is called a **strictly unimodal function** over a real domain $x \in \mathbf{D}$ if there exists some value $m \in \mathbf{D}$ such that f is strictly increasing $\forall x \leq m$ and f is strictly decreasing $\forall x \geq m$.*

When reasoning about how two agents can jointly reach an agreement, it is useful to start by considering the scenario where one player is *all-powerful*, meaning the bargain is determined solely to maximize one player's utility. The formal definition of this sort of bargaining arrangement is provided below.

Definition 2.4 (Powerful-P solution). *For a given fine-tuning game player $P \in \{G, D\}$, the powerful-P solution is the revenue-sharing agreement $\delta^{Powerful\ P} \in [0, 1]$ that maximizes P's utility:*

$$\delta^{Powerful\ P} = argmax_{\delta \in [0,1]} U_P(\delta).$$

2.3 Focus on Unimodal Utilities

We are now in a position to state our first theorem, which characterizes the Pareto-optimal solutions to any fine-tuning game with strictly unimodal utility functions.

Theorem 2.1. *Consider a fine-tuning game where players bargain over a parameter δ. If the players' utilities are strictly unimodal functions of δ, the set of Pareto-optimal agreements is the interval between their optima $\{\delta^{Powerful\ D}, \delta^{Powerful\ G}\}$, where both players' utilities are greater than the disagreement scenario. If no such interval exists, then disagreement is Pareto-optimal.*

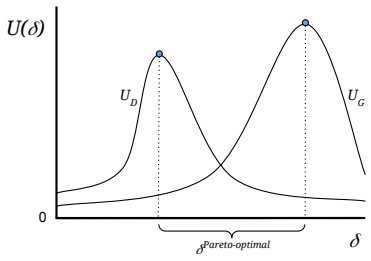

Figure 2: Example to illustrate Theorem 2.1. For two strictly unimodal, positive utility functions over bargaining parameter δ, the Pareto set is the interval between their optima.

The proof is provided in Appendix 7. To provide some intuition for the proof, consider the range of agreements δ between the point which maximizes one player's utility (say, $\delta^{Powerful\ D}$) and the point which maximizes the other ($\delta^{Powerful\ G}$). Agreements within this range exhibit a trade-off between the two utilities. Agreements outside this range, however, leave both players worse-off than, e.g., the nearest powerful-P solution, so they are Pareto-dominated. This intuition is illustrated in Figure 2.

Theorem 2.1 applies to a notably broad set of utility functions. For example, any strictly increasing, strictly decreasing, or strictly concave function on the interval $\delta \in [0, 1]$ is also strictly unimodal.

Equipped with the theorem above, solving the fine-tuning game consists of the following steps: (**1**) Use backward induction to solve for D and G's strategies, represented by α_1^* and α_0^*, in terms of δ. (**2**) Find the set of Pareto-optimal bargaining agreements δ between the powerful-D and powerful-G solutions. (**3**) Within the Pareto set, solve for bargaining agreements that maximize some joint function of the players' utilities.

3 ANALYSIS FOR POLYNOMIAL COSTS

Our model applies to general cost and revenue functions, and in Section 4 we provide results at this general level. But to understand how the central parameters of the model interact in closed form, it is also useful to study instantiations of the model with specific functional forms. Accordingly, we show in this section how to solve the model with a set of polynomial cost functions as a paradigmatic instance of convex cost functions, where the marginal costs increase as the technology is improved. Following this, we show how to draw conclusions about the model with general costs. In this section, cost functions take the following polynomial function forms:

$$\phi_0(\alpha_0) := c_0 \alpha_0^{k_0}, \;\; \phi_1(\alpha_1; \alpha_0) := c_1(\alpha_1 - \alpha_0)^{k_1}. \tag{2}$$

Here, $c_0, c_1 > 0$ since costs should increase with investment, and $k_0, k_1 > 1$, meaning that an incremental improvement grows costlier at higher levels of performance. We will continue to assume that $r(\alpha_1) = \alpha_1$ throughout this section's analysis.

First (3.1), we derive the subgame perfect equilibrium strategies α_0^*, α_1^* for fixed δ. Second (3.2), we find the set of Pareto-optimal revenue-sharing schemes δ^{Pareto}. Reaching a revenue-sharing agreement $\delta^* \in \delta^{Pareto}$ is modeled as a bargaining problem because the players must decide how to share surplus utility. So, third (3.3), we define five potential bargaining solutions: Best-performing-model, Vertical Monopoly, Egalitarian, Nash Bargaining

Solution, and Kalai-Smorodinsky. Where possible, we derive closed-form expressions for these solutions. We end by discussing the implications of these different revenue-sharing schemes.

3.1 Subgame Perfect Equilibrium for a Given δ

We use backward induction to determine the fine-tuning game's subgame perfect equilibrium (which we will refer to as a 'solution' or 'equilibrium'). Fixing the outcome of the initial negotiation, δ, it is possible establish the following closed-form solution:

Theorem 3.1. *For a fixed δ, the sub-game perfect equilibrium of the fine-tuning game with polynomial costs yields the following best-response strategies:*

$$\alpha_0^* = \left(\frac{\delta}{k_0 c_0}\right)^{\frac{1}{k_0-1}}, \quad \alpha_1^* = \left(\frac{\delta}{k_0 c_0}\right)^{\frac{1}{k_0-1}} + \left(\frac{1-\delta}{k_1 c_1}\right)^{\frac{1}{k_1-1}}.$$

A proof of the above result is provided in Appendix 8. Notice that the domain-specific performance, α_1^*, is equal to the general-purpose performance, α_0^*, plus a term, $(\frac{1-\delta}{k_1 c_1})^{\frac{1}{k_1-1}}$, independent of the G's choice over α_0^*. This is because the cost of marginal improvements for D only depends on the *difference* $(\alpha_1 - \alpha_0)$, and is not affected by a large or small initial investment by G. Though we assume, in this section, that D's cost is defined solely in terms of marginal improvement, Section 4 contains findings that generalize beyond this assumption, and further results are provided in the full version of this paper [34].

In order to determine the set of Pareto-optimal agreements, we first find that the utility functions are strictly unimodal functions of δ for all c_0, c_1 and $k_0, k_1 \geq 2$.

Proposition 3.1. *In the fine-tuning game with polynomial costs, if $k_0, k_1 \geq 2$, then U_G, U_D are strictly unimodal functions of $\delta \in [0, 1]$.*

The proof for the finding above, as well as all subsequent stated results other than theorems, is available in the full version of this paper [34]. The above finding suggest that the family of polynomial cost functions yield strictly unimodal utility curves. The set of Pareto-optimal solutions to these games can therefore be identified using Theorem 2.1. It is easy to show that the strict unimodality finding further generalizes to linear combinations of polynomial terms of the form provided in Equations (2), so long as all exponents are greater than or equal to 2. However, when the condition is not met and $k_0, k_1 < 2$, numerical simulations suggest that there are counter-examples to the strict unimodality property. When the strict unimodality property does not hold, it is still possible to analyze players' strategies—for example, our analysis in Section 4 stands even in cases where utility functions are not unimodal in δ.

Solving the powerful-G, powerful-D, vertical monopoly or other bargaining solutions consists in maximizing players' utilities either separately or combined into a joint utility. This is possible once parameters are specified; however, we cannot produce a closed-form expression for the general polynomial case because doing so would require solving for the zeroes of a polynomial of high degree. Therefore, for the remainder of this section, we will demonstrate the solution steps using parameter values $k_0, k_1 = 2$. We call this the case of *quadratic costs*. We choose the quadratic case for clarity and exposition, though we note that other solutions with other parameter values can be calculated using analogous steps.

3.2 Pareto-optimal Agreements on δ

We've derived both players' optimal strategies for fixed δ. Now, we consider the process where players agree on a particular value of δ. Since both players must enter an agreement in order for the technology to be viable, the determination of δ is a two-player bargaining game. We start by solving for the set of Pareto-optimal bargaining agreements, which is the interval between the 'powerful-player' solutions, defined below.

3.2.1 Powerful-Player Solutions. As we showed in Theorem 2.1, identifying the 'powerful-player' agreements is important for characterizing the set of Pareto-optimal bargaining solutions. Thus, we begin this section of analysis by solving for the powerful-G and powerful-D solutions (as defined in Definition 2.4).

Proposition 3.2 (Powerful-G Solution). *The Powerful-G solution to the model fine-tuning game with quadratic costs is as follows:*

$$\delta^{Powerful\ G} = \begin{cases} \frac{c_0}{2c_0 - c_1} & for\ c_1 < c_0, \\ 1 & for\ c_1 \geq c_0. \end{cases}$$

Proposition 3.3 (Powerful-D Solution). *The Powerful-D solution to the model fine-tuning game with quadratic costs is as follows:*

$$\delta^{Powerful\ D} = \begin{cases} 0 & for\ c_1 < c_0, \\ \frac{c_1 - c_0}{2c_1 - c_0} & for\ c_1 \geq c_0. \end{cases}$$

Now, using Theorem 2.1 and Proposition 3.1, we can define the set of Pareto-optimal solutions as: $\delta^{Pareto} \in \{\delta : \delta \leq \delta^{Powerful\ G} \cap \delta \geq \delta^{Powerful\ D}\}$. A visual representation of these solutions for the fine-tuning game with quadratic costs is provided in Figure 4.

3.3 Bargaining Solutions to Specify δ

If neither player dominates in a bargain, how do they decide how to share surplus profit? Solutions to bargaining problems identify an agreement that maximizes some joint utility function or satisfies certain desirable properties. In this section, we define the various bargaining solutions that the two players could plausibly arrive at within the set of Pareto-optimal solutions. These solutions mostly use a joint utility function to guide the bargaining agreement, as depicted in Figure 3. A visual representation of the bargaining solutions is provided in Figure 4. Definitions and closed-form solutions are provided below, and the proofs and steps yielding the solutions are included in Appendix 8.

Solution that maximizes the technology's performance. The first solution we propose presumes the joint goal of the two players is to collectively produce a technology with maximum performance α_1^*. There are a few ways to think of this quantity: It is the performance of the technology, and, equivalently, it is also the amount of revenue the two players collect. Though we do not formally specify a social welfare function, the technological performance can be thought of as the total utility offered to society by firms G and D.

Definition 3.1 (Maximum-performance solution). *For the fine-tuning game, the maximum-performance bargaining solution is the feasible revenue-sharing agreement $\delta^{max\text{-}\alpha_1^*} \in [0, 1]$ that maximizes the technology's performance α_1^*: $\delta^{max\text{-}\alpha_1^*} = argmax_{\delta \in [0,1]} \alpha_1^*$.*

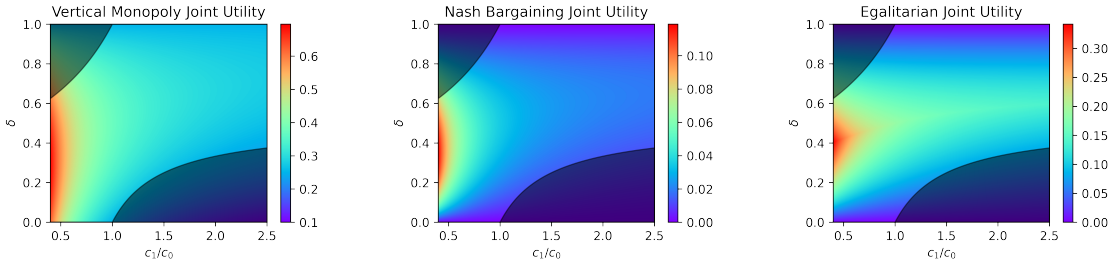

Figure 3: Various joint-utility functions for finding bargaining solutions. Gray regions are δ values that are not Pareto-optimal and therefore not candidate bargaining solutions. Color bar scales are defined assuming $c_0 = 1$.

Proposition 3.4 (Maximum-α_1^* Solution). *A bargaining solution that maximizes the technology's performance is given by:*

$$\delta^{\text{Max-}\alpha_1^*} = \begin{cases} 0 & \text{for } c_1 < c_0, \\ 1 & \text{for } c_1 \geq c_0. \end{cases}$$

Vertical Monopoly Solution. A perhaps intuitive approach to bargaining is to choose a revenue-sharing agreement that maximizes the sum of utilities $U_G + U_D$. This solution imagines that the two players are jointly controlled by a single entity who simply wishes to maximize the sum of utility. This solution is known as either the 'vertical monopoly' solution or the 'utilitarian' solution.

Definition 3.2 (Vertical Monopoly Solution). *For the fine-tuning game, the Vertical Monopoly (or 'Utilitarian') Solution is the feasible revenue-sharing agreement $\delta^{VM} \in [0,1]$ that maximizes the sum of the players' utilities: $\delta^{VM} = argmax_{\delta\in[0,1]}\,(U_G(\delta) + U_D(\delta))$.*

Proposition 3.5 (Vertical Monopoly Solution). *The Vertical Monopoly Bargaining Solution to the fine-tuning game with quadratic costs is as follows:*

$$\delta^{\text{Vertical Monopoly}} = \frac{c_1}{c_1 + c_0}.$$

Egalitarian Bargaining Solution. An alternative bargaining approach tries to help the worst-off player. This bargaining solutions is known as the 'egalitarian' solution.

Definition 3.3 (Egalitarian Bargaining Solution). *For the fine-tuning game, the Egalitarian Bargaining Solution is the feasible agreement $\delta^{Egal.} \in [0,1]$ that maximizes the minimum of players' utilities: $\delta^{Egal.} = argmax_{\delta\in[0,1]}(min_{P\in\{G,D\}}\,(U_P(\delta)))$.*

Proposition 3.6 (Egalitarian Bargaining Solution to the fine-tuning game with quadratic costs). *The Egalitarian Bargaining Solution to the fine-tuning game with quadratic costs is:*

$$\delta^{Egal.} = \frac{-\sqrt{c_0^2 - c_0c_1 + c_1^2} - c_1 + 2c_0}{3(c_0 - c_1)}.$$

Nash Bargaining Solution. The Nash Bargaining solution maximizes the product between the two players' utilities. This arrangement satisfies a number of desiderata, originally laid out by [39].

Definition 3.4 (Nash Bargaining Solution). *For the fine-tuning game, the Nash Bargaining Solution is the feasible revenue-sharing agreement $\delta^{NBS} \in [0,1]$ that maximizes the product of the players' utilities: $\delta^{NBS} = argmax_{\delta\in[0,1]}\,(U_G(\delta) * U_D(\delta))$.*

Though a closed-form solution for quadratic functions is possible, it involves solving the roots of a cubic function and yields a solution that is clunky and uninterpretable. We refer the reader to our numerical findings on this solution, depicted in Figures 3 and 4.

Kalai-Smorodinsky Bargaining Solution. Another solution suggested in economic literature, known as the Kalai-Smorodinsky bargaining solution, equalizes the ratio of maximal gains. Formally:

Definition 3.5 (Kalai-Smorodinsky Bargaining Solution [31]). *For the fine-tuning game, the Kalai-Smorodinsky Bargaining Solution (KSBS) is the feasible revenue-sharing agreement $\delta^{KSBS} \in [0,1]$ that satisfies the following relation:*

$$\frac{U_G(\delta^{KSBS})}{\max_{\delta\in\delta^{Pareto}} U_G(\delta)} = \frac{U_D(\delta^{KSBS})}{\max_{\delta\in\delta^{Pareto}} U_D(\delta)}.$$

Notice the denominators in the above equation are simply the utilities associated with the powerful-G and powerful-D solutions. Despite this simplifying step, the closed form Kalai-Smorodinsky solution is clunky and uninterpretable, so we omit it from this paper. Our numerical findings on this solution are depicted in Figure 4.

3.4 Discussion on Bargaining Solutions

Above we solve for a number of bargaining solutions revealing different possible configurations of fine-tuning arrangements. The general technology-producer and the domain specialist each have different optimal arrangements, between which any agreement is Pareto-optimal in the case of polynomial costs.

The first notable take-away is that players do not necessarily opt to maximize their own proportion of the profit. Even if one player has full control over the bargaining solution, depending on the relative cost of production, they may benefit from a profit-sharing agreement in order to encourage investment by the other player. If bargaining is conceptualized as splitting a pie, one player prefers to cede some portion of the pie if it means the entire pie grows to a size that justifies profit-sharing. This phenomenon arises in real-world settings. For instance, Apple allows third party developers to build software on iPhones. Opening up the tasks of application development to third parties improves consumer experience such that consumers are willing to purchase apps or other capabilities within apps. This additional revenue is then shared between Apple and the developer, leaving Apple with higher profits and a better product. Revenue sharing arises, often, because doing so is lucrative.

Profit-sharing is present even when both players have exceedingly different costs of production (i.e., when c_1/c_0 approaches 0

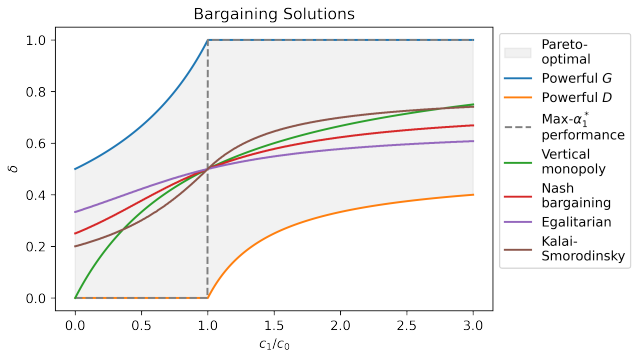

Figure 4: Bargaining agreements for the fine-tuning game with quadratic costs. Most bargaining solutions involve revenue sharing, even when one player faces much higher costs.

or ∞). In these limiting instances, we find that the Nash bargaining solution, Kalai-Smorodinski, and Egalitarian solutions all suggest profit-sharing. Only the Utilitarian solution—which models the two players as a vertical monopoly that is centrally controlled—yields the intuitively performance-optimal bargain, where the player with lower costs receives the entire profit. However, the vertical monopoly solution is not always performance-optimal. It underperforms the KSBS when the players face similar costs ($\sim 0.5 < \frac{c_1}{c_0} < 2.5$).

The bargaining solutions are neither binding rules nor descriptive observations; instead, they can be thought of as *normative* prescriptions. Identifying joint utility functions can help guide agents towards decisions that serve collective interests. For example, utilitarian and egalitarian solutions offer different visions for the appropriate distribution of welfare. In the same vein, one could specify and commit to a *social welfare function* in order to identify a bargaining solution that might be referred to as 'socially optimal.' Unsurprisingly, however, specifying social interests in a single function is an ambitious undertaking. In our present case, a social welfare function would need to balance the interests of (at least) 1) the technology's producers 2) consumers who value performance and 3) other external stakeholders. The procedure demonstrated in this section provides a road map for a social welfare analysis of the deployment of general-purpose models. Such an analysis might uncover how fine-tuning processes can be configured to serve collective, societal interests.

4 MULTIPLE DOMAIN SPECIALISTS

So far, we have modeled the fine-tuning process as a two-player game between a generalist and a single specialist. However, an important feature of general-purpose AI models is that they can be developed without fully anticipating the set of possible downstream use-cases. To capture the possibly many use-cases for general-purpose models, in this section, we generalize our model to the case where $n \geq 1$ domain specialists adapt the technology.

The multi-specialist fine-tuning game. Consider a game with $n \geq 1$ specialists. The players are G, D_1, D_2, ... D_n and we use i to index the specialists. G develops a technology to general performance α_0, after which every domain specialist D_i invests in the technology, bringing it to performance α_i in their domain. G and D_i share revenue $r_i(\alpha_i)$ according to bargaining parameter $\delta_i \in [0, 1]$. At the end of the game, G receives $\sum_i \delta_i r_i(\alpha_i)$ and each specialist D_i receives $(1 - \delta_i) r_i(\alpha_i)$. The game involves the following steps:

(1) Players bargain to decide δ_i for every domain i.
(2) G invests in a general-purpose technology yielding performance level α_0 and subject to cost $\phi_0(\alpha_0)$.
(3) Each specialist D_i may fine-tune the technology by choosing a performance level α_i subject to cost $\phi_i(\alpha_i; \alpha_0)$.

Players' utilities are defined as their revenue share minus cost:

$$U_G(\delta) := \sum_i \delta_i r_i(\alpha_i) - \phi_0(\alpha_0), U_{D_i}(\delta) := (1-\delta_i) r_i(\alpha_i) - \phi_i(\alpha_i; \alpha_0).$$

If G does not agree to a feasible bargain, she can instead opt for *disagreement*, where G receives utility d_0 and every specialist D_i receives utility d_i. If any particular domain specialist D_i does not agree to a feasible bargain, they may opt to receive d_i. However, this does not preclude other specialists from reaching a deal or adapting the technology. We assume, unless otherwise specified, that the disagreement scenario is described by $d_0 = d_i = 0$ for all i.

In the full version of the paper, we analyze all three steps of the game defined above, beginning with bargaining over δ_i [34]. Here, for brevity, we focus on step (3), in which the δ_i values have been determined and G has made an investment, and each specialist must now decide how much to spend on increasing the performance within their domain. In the full version [34], we also analyze a different form of the multi-specialist game where there is a single bargaining parameter shared by all domain specialists. This might describe, for example, app stores, where all applications follow the same revenue-sharing agreement [47].

4.1 Domain Specialists' Equilibrium Strategies

When there are potentially many domains where a technology may prove useful or marketable, different strategies around investment levels and fine-tuning can arise. In some domains, a technology may be adopted 'as-is' without significant additional investment or specialization. In other domains, it might be in everyone's interest for a technology to receive significant investment and specialization. Of course, in other domains, a technology might not be viable for any use at all. In this section, we explore the different sorts of cooperation (or non-cooperation) that can arise in domains with different characteristics. Our next general finding is a theorem on the various regimes of domain specialist strategies, depending on certain attributes of revenue and cost functions.

First, we will offer a set of relevant definitions to help characterize the different possible regimes of strategies for the specialist. Then, we will state the formal theorem.

Definition 4.1 (Contributor). *A domain specialist D_i is a **contributor** at the profit-sharing agreement δ_i if, given the generalist's optimal investment α_0 at δ_i, D_i's optimal strategy is to bring the technology to performance $\alpha_i^* > \alpha_0$.*

Definition 4.2 (Free-rider). *A domain specialist D_i is a **free-rider** at the profit-sharing agreement δ_i if, given the generalist's optimal investment α_0 at δ_i, D_i's optimal strategy is to enter the deal without improving the technology's performance, so $\alpha_i^* = \alpha_0$.*

Definition 4.3 (Abstainer). *A domain specialist D_i is an **abstainer** at the profit-sharing agreement δ if, given the generalist's*

optimal investment α_0 at δ, D_i's optimal strategy is to exit the deal and opt for disagreement.

Notice that any specialist is inevitably either a contributor, a free-rider, or an abstainer. These three regimes span the possible strategies for D_i. Below, we outline conditions that characterize D_i's strategy depending on their domain's cost and revenue $\{r_i, \phi_i\}$.

Theorem 4.1. *Suppose G has produced a general-purpose technology operating at performance α_0 and available at profit-sharing parameter δ_i. For any specialist with utility unimodal in α_i, the following conditions characterize their strategy, as shown in Table 1.*

- *"Fixed Costs Under Control" (FCUC): At zero investment ($\alpha_i = \alpha_0$), the domain specialist i's cost is less than its share of the revenue. Formally, $r_i(\alpha_0) > \frac{1}{1-\delta_i}\phi_i(\alpha_0; \alpha_0)$.*
- *"Marginally Profitable Investment" (MPI): At zero investment ($\alpha_i = \alpha_0$), a marginal investment from the domain specialist i increases its revenue share more than its costs. Formally, $r_i'(\alpha_0) > \frac{1}{1-\delta_i}\phi_i'(\alpha_0; \alpha_0)$.*

"FCUC"	*"MPI"*	*Type of Specialist*
T	*T*	*Contributor*
T	*F*	*Free-rider*
F	*T*	*Contributor or Abstainer**
F	*F*	*Abstainer*

Table 1: Types of specialists. In the third case (*), marginal conditions alone do not determine whether the specialist contributes or abstains.

A proof of the above theorem is provided in Appendix 9. The requirement that specialist utility is unimodal in α_i is, in our view, quite natural and broad. It covers three possible scenarios: 1) utility is increasing with investment, 2) utility is decreasing with investment, or 3) utility increases with investment up to a certain point, beyond which any further investment is not cost-justified.

It is important to note that the three regimes defined in this section can describe a specialist's strategy in either the 1-specialist or multi-specialist fine-tuning game. In the 1-specialist case, the potential strategies describe counterfactual outcomes that depend on the particular cost and revenue functions of the specialist. In the multi-specialist game, the strategies are ways of grouping the domains and all can exist simultaneously.

One scenario portrayed in Table 1 does not determine cleanly which regime the specialist falls into. In the scenario labeled with an asterisk (*), fixed costs are not under control but it is marginally profitable to invest in the technology. At zero investment, the technology is not ready to bring to market profitably, and it is unclear only from the marginal return on an initial investment whether it is worthwhile for the specialist to invest. In this case, the technology is potentially viable with some non-zero effort or, alternatively, not viable for the domain at any level of investment. Though the marginal conditions do not tell us whether the specialist would contribute or abstain, we can identify their strategy as follows: If $(1-\delta_i)r_i(\alpha_i) - \phi_i(\alpha_i)$ has positive real roots (for values of α_i greater than α_0), then D_i would contribute. Otherwise, D_i would abstain.

An illustration of Theorem 4.1 is provided in Figure 5. A noteworthy feature of this result is that it allows us to identify particular adaptation strategies using only the attributes about the domain i

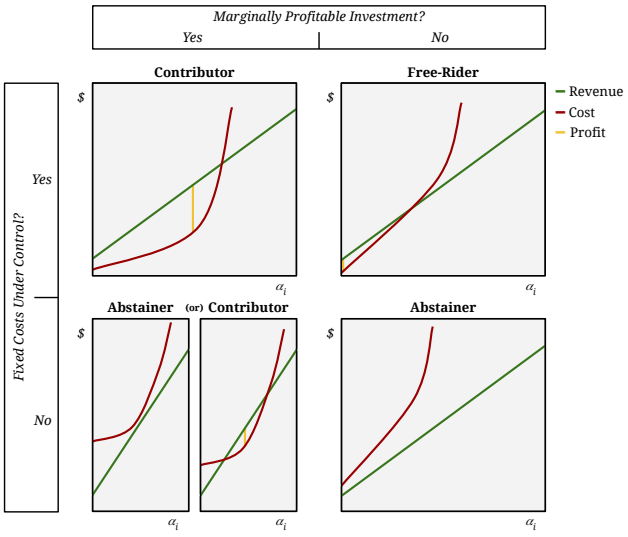

Figure 5: Examples illustrating Theorem 4.1. Depending on the characteristics of the cost and revenue curves, a domain specialist might engage in different types of strategy. For instance, when fixed costs are under control but investment is not marginally profitable (upper right quadrant), the firm will free-ride. When fixed costs are too high but investment yields marginal returns (lower left), the firm either abstains, or contributes if revenue exceeds cost at any point.

around $\alpha_i = \alpha_0$. In this setting, much of the information about the viability of a technology can be learned from only the 0th- and 1st-order approximations of $U_{D_i}\big|_{\alpha_i=\alpha_0}$, when the domain has invested minimal effort in the technology. This result perhaps coheres with the belief that a 'minimum viable product' (MVP) can provide an important signal about the profitability of a technology [3, 45].

Our analysis helps explain why technologies see significant uptake in some domains and not others. It characterizes domains that are particularly suitable or unsuitable to adopt a general-purpose technology. It also may explain why some technologies are re-sold without additional investment while others require fine-tuning.

5 CONCLUSION

Our model provides a starting point for considering the different interests and choices involved in the development of general-purpose models. By putting forward this model, we attempt to invoke the political economy of the development of AI technologies. These technologies are produced by a number of entities with different interests, and may potentially affect many individuals. This paper models agents' different interests explicitly, and proposes methods for weighing between them in light of societal values.

The work suggests a number of interesting directions for further research. One direction is to identify further general existence results for bargaining solutions with general functions in this model. More broadly, we also believe that formalizing the societal interests involved in AI regulation is an important direction. Such a formalism would need to build on an underlying model that contains the economic interests of the firms producing the AI technology. Our model may therefore help form the foundation for such work.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. ACM, Barcelona, Spain, 252–260.

[2] Robert J Aumann. 1961. The core of a cooperative game without side payments. *Trans. Amer. Math. Soc.* 98, 3 (1961), 539–552.

[3] Steve Blank. 2018. Why the lean start-up changes everything.

[4] 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. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021), 1–217.

[5] Rishi Bommasani, Dilara Soylu, Thomas I Liao, Kathleen A Creel, and Percy Liang. 2023. Ecosystem Graphs: The Social Footprint of Foundation Models. *arXiv preprint arXiv:2303.15772* (2023), 1–28.

[6] Timothy Bresnahan. 2010. General purpose technologies. *Handbook of the Economics of Innovation* 2 (2010), 761–791.

[7] Benedetta Brevini and Frank Pasquale. 2020. Revisiting the Black Box Society by rethinking the political economy of big data. , 205395172093514 6 pages.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona, Spain, 1186–1197.

[10] Robert G Cooper. 1984. The strategy-performance link in product innovation. *R&D Management* 14, 4 (1984), 247–259.

[11] Nicholas Crafts. 2021. Artificial intelligence as a general-purpose technology: an historical perspective. *Oxford Review of Economic Policy* 37, 3 (2021), 521–536.

[12] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. *Advances in neural information processing systems* 28 (2015), 1–10.

[13] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305* (2020), 1–11.

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

[15] Theo SH Driessen. 2013. *Cooperative games, solutions and applications*. Vol. 3. Springer Science & Business Media, Germany.

[16] Shaofu Du, Tengfei Nie, Chengbin Chu, and Yugang Yu. 2014. Newsvendor model for a dyadic supply chain with Nash bargaining fairness concerns. *International Journal of Production Research* 52, 17 (2014), 5070–5085.

[17] Francis Ysidro Edgeworth. 1881. *Mathematical psychics: An essay on the application of mathematics to the moral sciences*. C. Kegan Paul and Co., London, England.

[18] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130* (2023), 1–36.

[19] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications* 13, 1 (2022), 3094.

[20] Christian Fischer and Hans-Theo Normann. 2019. Collusion and bargaining in asymmetric Cournot duopoly—An experiment. *European Economic Review* 111 (2019), 360–379.

[21] Alfonso Gambardella and Anita M McGahan. 2010. Business-model innovation: General purpose technologies and their implications for industry structure. *Long range planning* 43, 2-3 (2010), 262–271.

[22] Avi Goldfarb, Bledi Taska, and Florenta Teodoridis. 2023. Could machine learning be a general purpose technology? a comparison of emerging technologies using data from online job postings. *Research Policy* 52, 1 (2023), 104653.

[23] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403* (2020), 1–15.

[24] Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. 2023. Algorithmic Collective Action in Machine Learning. *arXiv preprint arXiv:2302.04262* (2023), 1–21.

[25] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. Association for Computing Machinery, Cambridge, Massachusetts, 111–122.

[26] Oliver Hart and John Moore. 1988. Incomplete contracts and renegotiation. *Econometrica: Journal of the Econometric Society* 56, 4 (1988), 755–785.

[27] Elhanan Helpman. 1998. *General purpose technologies and economic growth*. MIT press, Cambridge, Massachusetts.

[28] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018), 1–12.

[29] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Atlanta, GA, 259–268.

[30] Boyan Jovanovic and Peter L Rousseau. 2005. General purpose technologies. In *Handbook of economic growth*. Vol. 1. Elsevier, Cambridge, MA, 1181–1224.

[31] Ehud Kalai and Meir Smorodinsky. 1975. Other solutions to Nash's bargaining problem. *Econometrica: Journal of the Econometric Society* 43, 3 (1975), 513–518.

[32] Jon Kleinberg and Manish Raghavan. 2020. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)* 8, 4 (2020), 1–23.

[33] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054* (2022), 1–54.

[34] Benjamin Laufer, Jon Kleinberg, and Hoda Heidari. 2023. Fine-Tuning Games: Bargaining and Adaptation for General-Purpose Models. *arXiv preprint arXiv:2308.04399* (2023), 1–36. https://arxiv.org/pdf/2308.04399.pdf

[35] Benjamin Laufer, Jon Kleinberg, Karen Levy, and Helen Nissenbaum. 2023. Strategic Evaluation. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Boston, MA, 1–12.

[36] Richard G Lipsey, Kenneth I Carlaw, and Clifford T Bekar. 2005. *Economic transformations: general purpose technologies and long-term economic growth*. Oxford University Press, Oxford, United Kingdom.

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

[38] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Atlanta, Georgia, 230–239.

[39] John F Nash. 1950. The bargaining problem. *Econometrica: Journal of the econometric society* 18, 2 (1950), 155–162.

[40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[41] Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, Virtual, 1314–1331.

[42] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, Online, 7599–7609.

[43] Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987* (2019), 1–8.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, Virtual, 8748–8763.

[45] Eric Ries. 2011. *The Lean Startup*. Crown Business, New York.

[46] William P Rogerson. 1992. Contractual solutions to the hold-up problem. *The Review of Economic Studies* 59, 4 (1992), 777–793.

[47] Paolo Roma and Daniele Ragaglia. 2016. Revenue models, in-app purchase, and the app performance: Evidence from Apple's App Store and Google Play. *Electronic commerce research and applications* 17 (2016), 173–190.

[48] Victor Sanh, Thomas Wolf, and Alexander Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems* 33 (2020), 20378–20389.

[49] Lloyd S Shapley. 1971. Cores of convex games. *International journal of game theory* 1 (1971), 11–26.

[50] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* 35, 5 (2016), 1299–1312.

[51] Ai Takeuchi, Róbert F Veszteg, Yoshio Kamijo, and Yukihiko Funaki. 2022. Bargaining over a jointly produced pie: The effect of the production function on bargaining outcomes. *Games and Economic Behavior* 134 (2022), 169–198.

[52] Manuel Trajtenberg. 2018. *AI as the next GPT: a Political-Economy Perspective*. Technical Report. National Bureau of Economic Research.

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

[54] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. 2017. Growing a brain: Fine-tuning by increasing model capacity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Honolulu, HI, USA, 2471–2480.

[55] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022), 1–30.

[56] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society* 10, 1 (2023), 20539517231177620.
[57] Desheng Wu, Opher Baron, and Oded Berman. 2009. Bargaining in competing supply chains with uncertainty. *European Journal of Operational Research* 197, 2 (2009), 548–556.
[58] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021), 1–17.
[59] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample BERT fine-tuning. *International Conference on Learning Representations* (2021), 1–22.
[60] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.

ACKNOWLEDGMENTS

The authors would like to thank the members of the AI, Policy and Practice group (AIPP) at Cornell University, the Center for Data Science (CDS) at NYU, the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) at Rutgers University, the Digital Life Initiative (DLI) at Cornell Tech, the EconCS group at Harvard University, and the Human + Machine Decisions Group at MIT for providing us the opportunity to present this work and for their feedback. In particular, we thank our anonymous reviewers, Sarah Cen, Yiling Chen, Nikhil Garg, James Grimmelmann, Karen Levy, Helen Nissenbaum, Kenny Peng, Manish Raghavan, Yoav Wald, and Lily Xu for illuminating conversations and suggestions.

The work is supported in part by a grant from the John D. and Catherine T. MacArthur Foundation. Ben Laufer is additionally supported by a LinkedIn-Bowers CIS PhD Fellowship, a doctoral fellowship from DLI, and a SaTC NSF grant CNS-1704527. Jon Kleinberg is additionally supported by a Vannevar Bush Faculty Fellowship, AFOSR award 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 are those of the authors and do not reflect the views of NSF or other funding agencies.

6 OVERVIEW OF DEFERRED PROOFS

Proofs for all theorems contained in this paper are provided in this appendix. Other stated results (e.g., propositions) are proven in the full version of this paper. See Laufer et al. [34].

7 SECTION 2 MATERIALS

Pareto set characterization and Theorem 2.1

Proof of Theorem 2.1. Consider three non-overlapping intervals that collectively span the feasible set $\delta \in [0, 1]$. These intervals are:

(1) $0 \leq \delta < \min(\delta^{\text{Powerful } D}, \delta^{\text{Powerful } G})$
(2) $\min(\delta^{\text{Powerful } D}, \delta^{\text{Powerful } G}) \leq \delta \leq \max(\delta^{\text{Powerful } D}, \delta^{\text{Powerful } G})$
(3) $\max(\delta^{\text{Powerful } D}, \delta^{\text{Powerful } G}) < \delta \leq 1$

We will characterize each of these intervals in turn, finding that intervals (1) and (3) are always Pareto dominated, and interval (2) is characterized by a trade-off in utilities.

(1) Within interval (1), the domain is characterized by $\delta < \min(\delta^{\text{Powerful } D}, \delta^{\text{Powerful } G}) \implies \delta < \delta^{\text{Powerful } D}$ and $\delta < \delta^{\text{Powerful } G}$. By the definition of a strictly unimodal function (2.3), this means that both utility functions $\{U_D, U_G\}$ are strictly increasing over interval 1. Thus, there exists some quantity $\epsilon > 0$ such that, for any value δ in interval (1), $U_D(\delta + \epsilon) > U_D(\delta)$ and $U_G(\delta + \epsilon) > U_G(\delta)$. Thus, every potential agreement in interval (1) is Pareto-dominated.
(2) Within interval (2), the domain is characterized by $\min(\delta^{\text{Powerful } D}, \delta^{\text{Powerful } G}) \leq \delta$, and also $\delta \leq \max(\delta^{\text{Powerful } D}, \delta^{\text{Powerful } G})$. If $\delta^{\text{Powerful } D} = \delta^{\text{Powerful } G}$, then the value $\delta = \delta^{\text{Powerful } D} = \delta^{\text{Powerful } G}$ is the unique Pareto-optimal agreement because it is optimal for both players. Otherwise if $\delta^{\text{Powerful } D} \neq \delta^{\text{Powerful } G}$, then interval (2) can be characterized as follows: For one player $P \in \{G, D\}$, the utility U_P one utility function is strictly decreasing because $\delta \geq \delta^{\text{Powerful } P}$ and $U_P(\delta)$ is a strictly unimodal function. For the other player $\{G, D\} \setminus P$, the utility $U_{\{G,D\}\setminus P}$ is strictly increasing because $\delta \leq \delta^{\text{Powerful } \{G,D\}\setminus P}$ and $U_{\{G,D\}\setminus P}(\delta)$ is a strictly unimodal function. Since one player's utility is strictly increasing and the other's is strictly decreasing, any perturbation of δ within interval (2) constitutes a utility gain for one player and a utility loss for the other. For any value of δ within this interval, if both players' utilities exceed the disagreement payoff (i.e., positive utility), then δ is Pareto-optimal.
(3) Within interval (3), the domain is characterized by $\delta > \max(\delta^{\text{Powerful } D}, \delta^{\text{Powerful } G}) \implies \delta > \delta^{\text{Powerful } D}$ and $\delta > \delta^{\text{Powerful } G}$. By the definition of a strictly unimodal function (2.3), this means that both utility functions $\{U_D, U_G\}$ are strictly decreasing over interval (3). Thus, there exists some quantity $\epsilon > 0$ such that, for any value δ in interval (3), $U_D(\delta - \epsilon) > U_D(\delta)$ and $U_G(\delta - \epsilon) > U_G(\delta)$. Thus, every potential agreement in interval (3) is Pareto-dominated.

Thus interval (2) is Pareto-efficient among the set of feasible bargaining agreements. □

8 SECTION 3 MATERIALS

Subgame perfect equilibrium findings

Proof of Theorem 3.1. We solve the game using backward induction as follows:

First, starting with the last stage (3), we solve for α_1^* given α_0, δ, c_1:

$$\begin{aligned}
& \alpha_1^* = \text{argmax}_{\alpha_1} U_D(\alpha_1, \alpha, \delta) \\
\Rightarrow \quad & \left.\frac{\partial U_D}{\partial \alpha_1}\right|_{\alpha_1=\alpha_1^*} = 0 \\
\Rightarrow \quad & \left.\frac{\partial}{\partial \alpha_1}\left((1-\delta)\alpha_1 - c_1(\alpha_1-\alpha_0)^{k_1}\right)\right|_{\alpha_1=\alpha_1^*} = 0 \\
\Rightarrow \quad & (1-\delta) - k_1 c_1 (\alpha_1^* - \alpha_0)^{k_1-1} = 0 \\
\Rightarrow \quad & \alpha_1^* = \alpha_0 + \left(\frac{1-\delta}{k_1 c_1}\right)^{\frac{1}{k_1-1}}. \qquad (3)
\end{aligned}$$

Note that $\frac{\partial^2 U_D}{\partial \alpha_1^2} = -k_1(k_1-1)c_1(\alpha_1-\alpha_0)^{k_1-2}$. This quantity is negative as long as $k > 1$, which is assumed. Thus, the α_1^* derived above yields a global maximum of U_D.

Second, knowing D's choice of α_1^* above, we solve for α_0^* as follows:

$$\begin{aligned}
& \alpha_0^* = \text{argmax}_{\alpha_0} U_G(\alpha_0, \delta) \\
\Rightarrow \quad & \left.\frac{\partial U_G}{\partial \alpha_0}\right|_{\alpha_0=\alpha_0^*} = 0 \\
\Rightarrow \quad & \left.\frac{\partial}{\partial \alpha_0}\left(\delta\alpha_1^* - c_0\alpha_0^{k_0}\right)\right|_{\alpha_0=\alpha_0^*} = 0 \\
\Rightarrow \quad & \left.\frac{\partial}{\partial \alpha_0}\left(\delta\left(\alpha_0 + \left(\frac{1-\delta}{k_1 c_1}\right)^{\frac{1}{k_1-1}}\right) - c_0\alpha_0^{k_0}\right)\right|_{\alpha_0=\alpha_0^*} = 0 \\
\Rightarrow \quad & \left.\frac{\partial}{\partial \alpha_0}\left(\delta\alpha_0 + [\text{const}] - c_0\alpha_0^{k_0}\right)\right|_{\alpha_0=\alpha_0^*} = 0 \\
\Rightarrow \quad & \delta - k_0 c_0 (\alpha_0^*)^{k_0-1} = 0 \\
\Rightarrow \quad & \alpha_0^* = \left(\frac{\delta}{k_0 c_0}\right)^{\frac{1}{k_0-1}}.
\end{aligned}$$

The second derivative $\frac{\partial^2 U_G}{\partial \alpha^2} = -k_0(k_0-1)c_0(\alpha_0)^{k_0-2}$. This quantity is negative as long as $k > 1$, which is assumed. Thus, the value of α_0^* derived above yields a global maximum of U_G.

Finally, plugging in $\alpha_0^* = \left(\frac{\delta}{k_0 c_0}\right)^{\frac{1}{k_0-1}}$ into Equation 3, we obtain the following expression for α_1^* as a function of δ only:

$$\alpha_1^* = \left(\frac{\delta}{k_0 c_0}\right)^{\frac{1}{k_0-1}} + \left(\frac{1-\delta}{k_1 c_1}\right)^{\frac{1}{k_1-1}}.$$

This finishes the proof. □

9 SECTION 4 MATERIALS

Theorem on the three specialist regimes

Proof of Theorem 4.1. We prove this theorem in a sequence of Lemmas. The proof follows for any given specialist D_i and revenue-sharing parameter δ_i.

Lemma 9.1. *If fixed costs are under control, meaning $r_i(\alpha_0) > \frac{1}{1-\delta_i}\phi_i(\alpha_0)$, then D_i will not abstain – instead, D_i would always prefer to free-ride.*

If $r_i(\alpha_0) > \frac{1}{1-\delta_i}\phi_i(\alpha_0)$, then $U_{D_i}\big|_{\alpha_i=\alpha_0} = r_i(\alpha_0) - \frac{1}{1-\delta_i}\phi_i(\alpha_0)$ is simply the RHS minus the LHS of the inequality. This means U_{D_i} must be positive at $\alpha_i = \alpha_0$. Thus, as long as fixed costs are under control, the specialist prefers free-riding to abstaining.

Lemma 9.2. *If fixed costs are not under control, meaning $r_i(\alpha_0) < \frac{1}{1-\delta_i}\phi_i(\alpha_0)$, then D_i will not free-ride – instead, D_i would always prefer to abstain.*

If $r_i(\alpha_0) < \frac{1}{1-\delta_i}\phi_i(\alpha_0)$, then $U_{D_i}\big|_{\alpha_i=\alpha_0} = r_i(\alpha_0) - \frac{1}{1-\delta_i}\phi_i(\alpha_0)$ is simply the RHS minus the LHS of the inequality. This means U_{D_i} must be negative at $\alpha_i = \alpha_0$. Thus, as long as fixed costs are not under control, the specialist prefers abstaining to free-riding.

Lemma 9.3. *If it is marginally profitable to invest in the technology, meaning $r_i'(\alpha_0) > \frac{1}{1-\delta_i}\phi_i'(\alpha_0)$, then D_i will not free-ride – instead, D_i would always prefer to contribute.*

If $r_i'(\alpha_0) > \frac{1}{1-\delta_i}\phi_i'(\alpha_0)$, then $\frac{\partial U_{D_i}}{\partial \alpha_i}\big|_{\alpha_i=\alpha_0} = r_i'(\alpha_0) - \frac{1}{1-\delta_i}\phi_i'(\alpha_0)$ is simply the RHS minus the LHS of the inequality. This means U_{D_i} is increasing at $\alpha_i = \alpha_0$. Thus, as long as it is marginally profitable to improve the technology, the specialist prefers contributing to free-riding.

Lemma 9.4. *If it is marginally costly to invest in the technology, meaning $r_i'(\alpha_0) < \frac{1}{1-\delta_i}\phi_i'(\alpha_0)$, then D_i will not contribute – instead, D_i would always prefer to free-ride.*

If $r_i'(\alpha_0) < \frac{1}{1-\delta_i}\phi_i'(\alpha_0)$, then $\frac{\partial U_{D_i}}{\partial \alpha_i}\big|_{\alpha_i=\alpha_0} = r_i'(\alpha_0) - \frac{1}{1-\delta_i}\phi_i'(\alpha_0)$ is simply the RHS minus the LHS of the inequality. This means U_{D_i} is decreasing at $\alpha_i = \alpha_0$. Thus, as long as it is marginally costly to improve the technology, the specialist prefers free-riding to contributing.

Taken together, we can conclude the following about combinations of conditions:

- Fixed costs under control, marginally profitable investment: A<F, F<C (Lemmas 9.1 and 9.3). Thus the specialist would contribute.
- Fixed costs under control, marginally costly: A<F, C<F (Lemmas 9.1 and 9.4). Thus the specialist would free-ride.
- Fixed costs not under control, marginally profitable: F<A, F<C (Lemmas 9.2 and 9.3). Thus the specialist would either abstain or contribute.
- Fixed costs not under control, marginally costly: F<A, C<F (Lemmas 9.2 and 9.4). Thus the specialist would abstain.

Above, the short-hand notation 'A,' 'F,' and 'C' refer to the strategies of abstaining, free-riding, and contributing, respectively. The optimal strategies follow from the two marginal conditions. This completes the proof. □