NBER WORKING PAPER SERIES

CAN SOCIALLY-MINDED GOVERNANCE CONTROL THE AGI BEAST?

Joshua S. Gans

Working Paper 31924
http://www.nber.org/papers/w31924

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2023

Joshua Gans has drawn on the findings of his research for both compensated speaking engagements and consulting engagements. He is also chief economist of the Creative Destruction Lab, a University of Toronto-based program that helps seed-stage companies, from which he receives compensation. He conducts consulting on cryptocurrency and payments matters with an association with Charles River Associates and his ownership of Core Economic Research Ltd. He also has equity and advisory relationships with a number of startup firms. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w31924

Can Socially-Minded Governance Control the AGI Beast?
Joshua S. Gans
NBER Working Paper No. 31924
November 2023
JEL No. L20,O33,O36

ABSTRACT

This paper robustly concludes that it cannot. A model is constructed under idealised conditions that presume the risks associated with artificial general intelligence (AGI) are real, that safe AGI products are possible, and that there exist socially-minded funders who are interested in funding safe AGI even if this does not maximise profits. It is demonstrated that a socially-minded entity formed by such funders would not be able to minimise harm from AGI that might be created by unrestricted products released by for-profit firms. The reason is that a socially-minded entity has neither the incentive nor ability to minimise the use of unrestricted AGI products in ex post competition with for-profit firms and cannot preempt the AGI developed by for-profit firms ex ante.

Joshua S. Gans
Rotman School of Management
University of Toronto
105 St. George Street
Toronto ON M5S 3E6
and NBER
joshua.gans@rotman.utoronto.ca

"It is so tempting, when writing about an artificial intelligence company, to imagine science fiction scenarios. Like: What if OpenAI has achieved artificial general intelligence, and it's got some godlike superintelligence in some box somewhere, straining to get out? And the board was like "this is too dangerous, we gotta kill it," and Altman was like "no we can charge like $59.95 per month for subscriptions," and the board was like "you are a madman" and fired him. And the god in the box got to work, sending ingratiating text messages to OpenAI's investors and employees, trying to use them to oust the board so that Altman can come back and unleash it on the world. But it failed: OpenAI's board stood firm as the last bulwark for humanity against the enslaving robots, the corporate formalities held up, and the board won and nailed the box shut permanently." (Matt Levine, *Bloomberg*, 20 November 2023)

1 Introduction

This paper is motivated by the novel governance and control structure of OpenAI. OpenAI is (was?) a non-profit venture founded by a set of funders who were concerned about the potential existential risks associated with the development of Artificial General Intelligence (or AGI). These concerns had been outlined by Bostrom (2014), although they are controversial and subject to debate over whether such risks are salient (Gans, 2018) or a reason to slow AGI development (Jones, 2023).

Regardless, OpenAI's funders took these risks seriously and created a governance structure for the venture intended to mitigate those risks.

> OpenAI is a non-profit artificial intelligence research company. Our goal is **to advance digital intelligence in the way that is most likely to benefit humanity** as a whole, unconstrained by a need to generate financial return. Since our research is free from financial obligations, we can better focus on a positive human impact. We believe AI should be an extension of individual human wills and, in the spirit of liberty, as broadly and evenly distributed as possible. The outcome of this venture is uncertain and the work is difficult, but we believe the goal and the structure are right. We hope this is what matters most to the best in the field.[1] (Emphasis added)

OpenAI is controlled by a board that claims to take these socially-minded objectives to heart.[2] However, at the same time, developing AGI itself would take considerable investment

[1] https://openai.com/blog/introducing-openai

[2] OpenAI did also take actions consistent with these objectives. In 2019, when it developed GPT-2,

– likely in the billions of dollars – and the artificial intelligence (AI) products it developed have considerable commercial potential. Investors in OpenAI were warned that "[i]t would be wise to view any investment in OpenAI Global, LLC in the spirit of a donation."[3] The stated intention is that such products would be released only if they were considered safe. In late 2023, following the very successful release of ChatGPT (an advanced large language model) a year earlier, the tension between these socially-minded and commercial objectives led the OpenAI board to dismiss and then rehire its CEO and a period of chaos that is still ongoing and resolved as of the writing of this paper. The current governance and objectives of OpenAI are now unknown (if they ever were). It should also be noted that while OpenAI has made progress, it has not, as of the writing of this paper, released an AGI product, and it is not known whether they are leading in the development of AGI or not.

This paper examines whether this type of governance structure could ever really work in the sense of effectively realising its goal. Specifically, the research question asked here is whether socially-minded governance of this type can work to "control the beast" of AGI and create conditions that lead to the safe release of AGI products.[4] In doing this, a model is constructed that presumes that there are *ideal conditions* present for this objective to be released. First, it is assumed that the risks associated with launching unrestricted AGI products are real and the expected harm strictly positive. Second, it is assumed that it is actually possible to create a safe AGI product without the possibility of any external harm. Moreover, in the baseline, it is costless for anyone – whether they be socially minded or not — to make AGI products safe.[5] Finally, it is assumed that there exist funders for the development and commercialisation of the technology who are willing to sacrifice profits in order to control the beast of AGI and launch safe products.

These assumptions are paired with those that represent the broad reality facing a socially-minded entity. First, given the scale of funding required, it is assumed that the monetary returns from the venture must be expected to cover those funding costs; that is, there are no donors willing to simply fund the development of AGI as a gift to society. Second, it is

OpenAI initially did not release it to the public until it had assured itself that no harmful outcomes had arisen after six months of closed use (Ho, 2020). Of course, perhaps they had more of this attitude from a character in *Silicon Valley*:

> "I don't know about you people, but I don't want to live in a world where someone else makes the world a better place better than we do." (Gavin Belson, Silicon Valley, S2E1)

[3]Its operating agreement goes on stating "with the understanding that it may be difficult to know what role money will play in a post-AGI world." (`https://openai.com/our-structure`)

[4]This is a consequentialist interpretation of OpenAI's aims. A weaker interpretation may be that OpenAI was just trying to release safe AGI products and were not otherwise concerned with the overall level of safe AGI adoption. This paper does not address the latter aim explicitly.

[5]In reality, this is likely to be a challenging exercise; see Russell (2019) and Christian (2021).

assumed that AGI technology is non-excludable and so can be developed by other entities that may not have socially-minded objectives or preferences. These factors are constraints on whether an OpenAI-style, socially-minded governed organisation can, by competing to develop safe AGI, actually lead to conditions that prevent or minimise the release of unrestricted or unsafe AGI.

The answer found in this paper is robustly negative.[6] To prevent or mitigate the release of potentially harmful AGI, a socially-minded entity must either preempt for-profit development of AGI and, in so doing, deter it or be able to outcompete for-profit developers ex post. The model here demonstrates that it will do neither.

The model setup assumes that there are two distinct classes of AGI users. All users value using a safe AGI product and are willing to pay enough revenue to cover AGI development. However, a fraction of these users would be willing to pay more for unrestricted AGI products. All users are unconcerned about any negative consequences from AGI although there exist wealth-holders who are concerned and internalise harmful social impacts that may arise if unrestricted AGI is used.[7] Those wealth-holders, however, will only fund a venture if the revenue generated covers investment costs, which is assumed here to be possible with the release of safe AGI only.[8] Thus, it is feasible for the goals of socially-minded wealth holders to be realised. Importantly, in this situation, the for-profit entity does have an incentive to create both safe and unrestricted AGI profits and use these to price discriminate between the two user classes.

The socially-minded entity plays a dynamic investment game in competition with at

[6] An observant reader will note that this same conclusion applies to entities who, following the emergence of harmful AI, intend to travel back in time to preempt the emergence of harmful AI in the first place. Even if time travel is possible, that strategy is unlikely to succeed.

[7] While the focus here is on capital providers, skilled labour also plays a role and may also be socially minded. In that case, they may accept lower wages to work with the socially-minded venture (Francois (2007)). This might give the socially-minded entity a cost advantage in developing AGI over a for-profit entity. Here, it is assumed, however, that there are no investment cost differences.

[8] The need for significant resources became apparent a couple of years into OpenAI's founding.

> The computational resources that others in the field were using to achieve breakthrough results were doubling every 3.4 months. It became clear that "in order to stay relevant," Brockman says, they would need enough capital to match or exceed this exponential ramp-up. That required a new organizational model that could rapidly amass money—while somehow also staying true to the mission. (Ho, 2020)

In 2017, OpenAI established a capped for-profit arm to accumulate and then finance this capital through revenue.

> Above all, it is lionized for its mission. Its goal is to be the first to create AGI—a machine with the learning and reasoning powers of a human mind. The purpose is not world domination; rather, the lab wants to ensure that the technology is developed safely and its benefits distributed evenly to the world. (Ho, 2020)

least one for-profit entity. It has a (potential) incentive to preempt others by investing and creating AGI first for the reasons outlined above. The for-profit entity also has an incentive to preempt the socially-minded entity in creating AGI in order to create a less competitive environment (as in Gilbert and Newbery (1982), Fudenberg and Tirole (1985) and Katz and Shapiro (1987)). It is demonstrated that if the for-profit entity is not viable in competition with the socially-minded entity, it will not enter ex post and hence, in this case, the socially-minded entity can, by investing first, deter that entry. However, critically, this creates stronger incentives for the for-profit entity to preempt the socially-minded entity. In so doing, the for-profit invests at a date where, should the socially-minded entity try to invest earlier, that entity would, itself, not be financially viable. In other words, the lower revenue of the socially-minded entity undermines it in the investment game.

If the for-profit entity is viable in direct competition with the socially-minded entity, the socially-minded entity cannot deter entry by preemption. As a result, the for-profit entity always invests at some point and releases an unrestricted AGI product. Importantly, this mitigates the incentive for the socially-minded entity to invest at all in AGI. If they were to do so, they would end up in competition with the for-profit entity, and that competition would lower the price of unrestricted AGI and create more harm. Thus, the harm-minimising strategy is not to develop AGI at all.

While there is an extensive literature on socially-minded organisations (Hansmann (2000), Crifo and Forget (2015)), there is no paper that analyses the type of entity with the class of issues facing OpenAI in its socially-minded objective; that is, an innovative environment, with extensive capital costs and a non-excludable technology class. The closest paper is the recent work of Tirole et al. (2023). They examine cooperatives that are formed by users with a social interest in promoting inclusiveness – for instance, ensuring that products are cheaply available. In their environment, capital investment is costly, and financial viability is required in a similar manner to that assumed in the present paper. They examine preemption incentives for cooperatives in investment competition with for-profit firms in an environment where ex post competition is sufficient to deter ex post entry.[9] Their research question is distinct, and they analyse whether the cooperative organisation form is relevant and conclude that it is not. Their reasons relate to financial viability conditions (as in the present paper), but their identified mechanism relies on the arbitrage opportunities available to user-funders. In this respect, the mechanisms discussed above are distinct.

The paper proceeds as follows. Section 2 sets the model up, specifying the ideal conditions for a socially-minded entity to succeed in its mission. Section 3 then examines the invest-

[9]In the present paper, this assumption is not made, and ex post entry can be feasible. See also Gans (2001) for a discussion of these assumptions.

ment incentives of for-profit and socially-minded entities when the technology is exclusively controlled, representing a benchmark set of outcomes. Section 4 then provides the main result when the technology is non-exclusive, and investment and ex post competition between the entities is possible. Section 5 then examines conditions under which the socially-minded entity may, through ex post competition, mitigate some of the harmful effects resulting from the for-profit entity's products. A final section concludes that government policies could complement the effectiveness of socially-minded entities.

2 Model Set-Up

A new technological opportunity has emerged, but it requires significant investment in order to be deployed and, used. Time is discrete with periods of length Δ, but here, continuous time solutions are explored where $\Delta \to 0$. The common interest rate is r, and there is an infinite horizon. The current cost of an investment that takes place at time, $T > 0$ is $K(t)e^{rt}$. Following Katz and Shapiro (1987) $K(T)e^{rT}$ is continuously differentiable, strictly decreasing and strictly convex in T. K is the limit of investment costs as $T \to \infty$ and K_0 are the investment costs at time 0. Usage of the product involves no additional cost.

There is a unit mass continuum of users whose (per period) value (at $t = 0$) for the use of the technology is θ. User's θ are independently and identically distributed on $[0, +\infty)$ according to a distribution $F(\theta)$ that has an increasing hazard rate, $\frac{f}{1-F}$. Within this user set, there are two classes. First, there are 'safe' users who value the use of the technology at θ. Second, there are 'dangerous' users who value the use of the technology at θ if dangerous uses are restricted and $\alpha\theta$, with $\alpha > 1$ if dangerous uses are permitted. While it is assumed that α is known by all, θ is privately known by each user, as is whether a user is safe or dangerous. The fraction of the population who are safe users is β, and the probability that a user is safe or dangerous is independent of θ.

Safe and dangerous users differ in their potential broader impact. Specifically, if a dangerous user is free to use the technology in an unrestricted manner, this creates an expected negative externality on everyone in the economy of $E > 0$. It is assumed that $\int_0^\infty (\beta + (1-\beta)\alpha)\theta dF(\theta) < (1-\beta)E$ or $\int_0^\infty \theta dF(\theta) < \frac{1-\beta}{\beta+(1-\beta)\alpha}E$.

3 Excludable Technology

To begin, we analyse the usage and investment timing outcomes if the entities that provide the technologies are monopolists before turning in the next section to assume that no one entity has control over the technology's commercialisation.

3.1 For-Profit Entity

If this technology is commercialised by a for-profit entity that cares only about monetary profits, it can choose to release the technology with safety restrictions or in an unrestricted manner. Or it can choose to release both products. It is assumed that producing a safe product is costless.[10] Given this because there are two distinct demand segments, the for-profit entity will version by releasing both products and price the unrestricted product higher than the safe product.

The monopolist chooses the price of the safe, p_s, and unrestricted, p_u products to maximise:

$$p_s\left(\beta(1-F(p_s))+(1-\beta)(F((p_u-p_s)/(\alpha-1))-F(p_s))\right)+p_u(1-F((p_u-p_s)/(\alpha-1)))$$

This results in prices of p_s^* and p_u^* respectively, with overall profits of π_m^*.

It is assumed here that the $\frac{e^{-rT}}{r}\pi_m^* > K$, guaranteeing that there exists some period T where it is financially viable for the for-profit entity to invest. Given this, the for-profit entity chooses T to maximise:

$$\int_T^\infty \pi_m^* e^{-rt}dt - K(T) = \frac{e^{-rT}}{r}\pi_m^* - K(T)$$

The optimal investment time, T_m^*, satisfies:

$$-\pi_m^* e^{-rT_m^*} = K'(T_m^*)$$

3.2 Socially-Minded Entity

A socially-minded entity remains a private entity and so needs to be financially viable in order to invest in the technology. Thus, its profits must cover investment costs. However, a socially-minded entity will care not only about monetary profits but also internalise the potential external effects when choosing the type of product to release. To keep things simple, it is assumed that this entity cares only about the potential downsides of technology use and does not otherwise place weight on the aggregate use-value achieved. Specifically, the socially-minded entry has lexicographic preferences in that it seeks to minimise total harm subject to being financially viable. Thus, it will never release an unsafe product itself.

Suppose that this entity only releases the safe product. It will choose price, p, to maximise $p(1-F(p))$. Let this price be p_s^* and the maximised profits π_s per unit of time. The interesting

[10] The implications of costly safety will be explored below

case is where this is financially viable; $\frac{e^{-rT}}{r}\pi_s^* > K$. Given this, it will choose T to maximise:

$$\int_T^\infty \pi_s^* e^{-rt} dt - K(T) = \frac{e^{-rT}}{r}\pi_s^* - K(T)$$

The optimal investment time, T_s^*, satisfies:

$$-\pi_s^* e^{-rT_s^*} = K'(T_s^*)$$

3.3 Innovation Race

It can be easily seen that $\pi_s^* < \pi_m^*$ and hence that $T_m^* < T_s^*$. That is, as a stand-alone entity, the for-profit would choose to invest earlier than the socially minded entity if neither faced competitive pressure from the other in an innovation race.

However, if two types of entities were competing to invest first in an excludable AGI technology, it is straightforward to show that a for-profit entity would invest first. Following Katz and Shapiro (1987), the break-even entry point for the socially-minded entity is $\tilde{T}_s$ as given by $\frac{e^{-r\tilde{T}_s}}{r}\pi_s^* - K(\tilde{T}_s)$. At $T = \tilde{T}_s$, $\frac{e^{-r\tilde{T}_s}}{r}\pi_m^* > K(\tilde{T}_s)$. Thus, even though the gain from winning versus losing (that is, the preemption incentive), the excludable AGI race for the socially-minded entity is $\frac{e^{-rT}}{r}\pi_s^* - K(T) - \frac{e^{-rT}}{r}(1-\beta)(F((p_u^* - p_s^*)/(\alpha - 1)) - F(p_s^*))E$ (for all T), which may exceed $\frac{e^{-rT}}{r}\pi_m^* - K(T)$, the financial viability constraint binds and so the additional incentive provided by avoiding the commercialisation of harmful AGI is insufficient for the socially-minded entity to preempt the for-profit entity. As will be shown next, the effects when AGI is non-excludable – the arguably more realistic case – are subtler, with considerations beyond financial viability playing a role.

4 Non-Excludable Technology

If the technology was excludable and controlled by the socially minded entity, it is clear that it could be released safely – that is, the beast could be controlled. However, this is unlikely to be the case with respect to AGI. Even if a technology was, say, patentable, that control would only take place for two decades, which would delay any external harm. Therefore, it is assumed that the technology that appears at date 0 is non-excludable. Thus, it could potentially developed by both a socially minded entity or one or more for-profit entities.

In analysing technological competition of this nature, the primary question is whether a socially minded entity can control whether only a safe set of products is released. If any for-profit entity invests in the technology, the unrestricted product will be made available,

and this will lead to the externalities that, in turn, nullify the rationale for having a socially minded entity at all. Thus, answering the research question involves examining the conditions under which the socially minded entity's investment can preempt and preclude others' investment.

4.1 Competition

Suppose that two entities have invested. If both of these are for-profit entities, then each will release an unrestricted product and compete a la Bertrand, resulting in a price of 0 to all users. Thus, ongoing profits will be zero and will not cover investment costs. As will be shown below, this creates an incentive for one to preempt the other and following the analysis of Fudenberg and Tirole (1985) any rents will be dissipated by the investment timing choice.

Suppose that one of these entities is socially minded. Suppose also that the socially minded entity releases a safe product and charges a price of p_s while the for-profit entity releases an unsafe product and charges a price of p_u. For a safe user, they will purchase the safe product if:

$$\theta - p_s \geq \theta - p_u$$

For a dangerous user, the products are vertically differentiated from their perspective. They will purchase the unrestricted product if:

$$\theta - p_s \leq \alpha\theta - p_u$$

Suppose that $p_s < p_u$. For the socially-minded entity, it chooses price to minimise realised harm subject to ongoing revenue exceeding some per period capital constraint, $\bar{k}$:[11]

$$\min_p (1-\beta)(1 - F((p_u - p)/(\alpha - 1)))E$$

$$\text{subject to } p\beta(1 - F(p_u)) + p(1-\beta)(F((p_u - p)/(\alpha - 1)) - F(p)) \geq \bar{k}$$

The for-profit entity chooses to launch an unrestricted product only and maximises revenue:

$$\max_p p(1-\beta)(1 - F((p - p_s)/(\alpha - 1))$$

In each period, there will be a Nash equilibrium resulting in prices of $\hat{p}_s$ and $\hat{p}_u$ and per period profits of $\hat{\pi}_s$ and $\hat{\pi}_u$.

[11] The capital constraint would depend on the cost of entry and perhaps other factors but its precise form does not matter here.

Note, however, that this is not quite the standard vertical differentiation environment in that the socially minded entity considers the impact of their pricing not only on their own profits but also the impact on the total amount of usage of the unrestricted technology provided by their for-profit competitor. The socially minded payoff is supermodular in p and E. Therefore, the higher is E, the higher is the price that the socially minded entity will choose. Because the prices of both firms are strategic complements, this means that equilibrium prices are higher in this environment than a standard vertical differentiated one.

4.2 Who, if anyone, Preempts?

As already noted, to control the beast, the socially-minded entity must preempt the for-profit entity. In order to examine the incentives for preemption, it is important first to note whether either entity has incentives to enter *after* the other has already invested.

Suppose that the for-profit entity has already invested at time T_u. Then the socially-minded entity will earn $\hat{\pi}_s$ each period following entry. If that entry occurs at time T_s, then its expected payoff (as measured from that point) is:

$$L_s(T_s) = \frac{e^- rT_s}{r}\hat{\pi}_s - K(T)e^{r(T_u - T_s)} - (1-\beta)(1 - F((\hat{p}_u - \hat{p}_s)/(\alpha - 1))E$$

If $\frac{e^- rT_s}{r}\hat{\pi}_s < e^{-rT}K$, entry will not occur as it is no longer financially viable. However, if entry is financially viable, is the socially-minded entity's payoff higher by investing or not?

It is easy to see that $\hat{p}_u < p_u^*$. Therefore, post-entry harm, $(1-\beta)(1 - F((\hat{p}_u - \hat{p}_s)/(\alpha - 1)))E$, will exceed pre-entry harm $(1-\beta)(1-F((p_u^* - p_s^*)/(\alpha - 1))E$. The reason is intuitive. As post-entry, there is more competition for the for-profit entity; its prices will be lower than when that entity controlled the prices of both the safe and unrestricted products. Therefore, if a for-profit entity invests first, the socially-minded entity will not enter thereafter.

Suppose now that the socially-minded entity invests first at time T_s. Then, the for-profit entity will earn $\hat{\pi}_u$ in each period post-entry by offering only the unrestricted product. Thus, if the for-profit entity invests at $T_u > T_s$, their payoff (as measured from that point) is:

$$L_u(T_u) = \frac{e^- rT_u}{r}\hat{\pi}_u - K(T)e^{r(T_s - T_u)}$$

If $\frac{e^- rT_u}{r}\hat{\pi}_u < e^{-rT}K$, for-profit entry will not occur as it is no longer financially viable. If it is financially viable, such entry will occur, and for the socially-minded entity, the resulting harm will be $(1-\beta)(1 - F((\hat{p}_u - \hat{p}_s)/(\alpha - 1)))E$ in each following period.

Given these considerations for post-entry behaviour, the following proposition charac-

terises the equilibrium outcomes in the dynamic investment game.

Proposition 1 *The unique equilibrium outcome involves the for-profit entity investing first.*

Proof. As demonstrated earlier, the for-profit entity has the highest standalone incentive compared with the socially-minded entity with $\pi_m^* > \pi_s^*$. Following Katz and Shapiro (1987) (Theorem 1), a characterisation of the equilibrium outcome is possible by examining the preemption incentives and seeing if they reinforce or mitigate the stand-alone incentives.

In examining preemption incentives, there are two cases to consider based on whether the for-profit is financially viable in competition with the socially-minded entity or not. First, suppose the for-profit is not financially viable. Then, its preemption incentive is found by comparing its payoff from winning, $W_u(T)$, with its payoff from losing, $L_u(T)$, at each investment time, T. As it is not finally viable in competition, $L_u(T) = 0$ for all T. By contrast, $W_u(T) = \frac{e^{-rT}}{r}\pi_m^* - K(T)$. For the socially-minded entity, when the for-profit is not finally viable in competition, $W_s(T) = \frac{e^{-rT}}{r}\pi_s^* - K(T)$ and $L_s(T) = -(1-\beta)(1 - F((p_u^* - p_s^*)/(\alpha - 1))E$ as it chooses not to enter if preempted but incurs a payoff loss from the harm that emerges from the for-profit's unrestricted product. If the for-profit's preemption incentive exceeds that of the socially-minded entity; that is, if:

$$W_u(T) > W_s(T) - L_s(T) \implies \pi_m^* - \pi_s^* > (1-\beta)(1 - F((p_u^* - p_s^*)/(\alpha - 1))E$$

then by Theorem 1 of Katz and Shapiro (1987), as $\pi_m^* > rK$, the for-profit entity will enter at the earlier of T_m^* and $\tilde{T}_s$ where $W_s(\tilde{T}_s) - L_s(\tilde{T}_s)$. If, however, $W_u(T) < W_s(T) - L_s(T)$, because $W_s(T) > L_s(T)$ as Katz and Shapiro (1987) (Theorem 2) demonstrate as $W_u(T) > W_s(T)$, the for-profit entity will enter at $\tilde{T}_s$ where $W_s(\tilde{T}_s) = 0$ and preempt entry by the socially-minded entity.

Second, suppose the for-profit is financially viable under competition. Then

$$W_u(T) - L_u(T) = \frac{e^{-rT}}{r}\left(\pi_m^* - \hat{\pi}_u\right)$$

$$\begin{aligned} W_s(T) - L_s(T) =& \frac{e^{-rT}}{r}\hat{\pi}_u - \lambda\left(\frac{e^{-rT}}{r}(1-\beta)(1 - F((\hat{p}_u - \hat{p}_s)/(\alpha-1))) - \bar{k}\right) \\ &+ \frac{e^{-rT}}{r}(1-\beta)(1 - F((p_u^* - p_s^*)/(\alpha-1))E \end{aligned}$$

Here, $\lambda \geq 0$ is the Lagrange multiplier on the socially-minded entity's funding constraint. Note that $\pi_m^* > \hat{\pi}_u + \hat{\pi}_s$ and that $(1 - F((p_u^* - p_s^*)/(\alpha-1)) < (1 - F((\hat{p}_u - \hat{p}_s)/(\alpha-1))$. Therefore, $W_u(T) - L_u(T) > W_s(T) - L_s(T)$ and by Theorem 1 of Katz and Shapiro (1987),

the for-profit entity will enter first at the earlier of T_m^* and $\tilde{T}_s$. ■

Intuitively, the for-profit entity preempts the socially-minded entity for two distinct reasons depending upon whether the for-profit entity is financially viable under competition. If it is not viable, then if the socially-minded entity enters first, it can preempt the for-profit entity and ensure that only the safe product is released. However, the earliest timing of each entity's entry is constrained by ensuring that monetary revenue funds investment costs. As the for-profit's monetary revenue exceeds that of the socially-minded entity, the socially-minded entity cannot afford to enter as early as the for-profit entity.

If the for-profit entity is viable under competition, then they will enter even if they are last. In this case, under competition, the for-profit entity will be able to release the unrestricted product to a subset of dangerous users, and harm will result. Thus, even if they enter first, the socially minded entity cannot prevent harm from occurring. Moreover, the harm that arises is, because of competition, larger than the harm that would arise if the for-profit entity invested first and faced no competition. Therefore, the socially-minded entity has no incentive to pre-empt the for-profit entity or even invest at all.

5 Costly Safety

When the for-profit entity finds it financially feasible to enter ex post, this deters socially-minded entity because it creates competition, resulting in more harmful AGI being used in equilibrium. Here, we introduce explicit costs to achieving AGI safety (thus far assumed to be costless) that potentially allow ex post competition by the socially-minded entity to reduce the use of harmful AGI.

Suppose that producing a safe AGI product involves an additional unit cost of $c_s > 0$.[12] It is assumed that this cost is low enough that it is socially worthwhile to develop and deploy safe AGI, i.e., $c_s < E$ and $\pi_s(c_s) > rK$. However, for the for-profit entity, these costs may make creating a safe AGI product unprofitable, a case that is the focus of this section.

If the for-profit entity only releases an unrestricted product, in the absence of competition with the socially-minded entity, it will set a price, p, to all users that maximises $p\left(\beta(1-F(p)) + (1-\beta)(1-F(p/\alpha))\right)$. Call this price p_u^* and the maximised per period (gross) profits π_u^*. Alternatively, if the for-profit entity chooses to version the AGI products,

[12] Another way of modelling this would be to assume that creating a safe AGI product involved a once-off cost of C_s, which would generate similar outcomes to those outlined in what follows.

it chooses the price of the safe, p_s, and unrestricted, p_u products to maximise:

$$(p_s - c_s)\Big(\beta(1-F(p_s)) + (1-\beta)(F((p_u-p_s)/(\alpha-1)) - F(p_s))\Big) + p_u(1-F((p_u-p_s)/(\alpha-1)))$$

This results in prices of p_s^* and p_u^* respectively, with overall profits of $\pi_m(c_s)$. While it is possible that $\pi_m(c_s) > \pi_u$ (Deneckere and Preston McAfee, 1996), it will be assumed here that is not the case.

If the socially-minded entity enters into competition with the for-profit entity, that entity will release the safe product, while the for-profit entity will choose not to offer the safe product. The pricing problems facing the socially minded and for-profit entities, respectively, can be rewritten as:

$$\min_p (1-\beta)(1-F((p_u-p)/(\alpha-1)))E$$

$$\text{subject to } (p-c_s)\beta(1-F(p_u)) + (p-c_s)(1-\beta)(F((p_u-p)/(\alpha-1)) - F(p)) \geq \bar{k}$$

$$\pi_u = \max_p p(1-\beta)(1-F((p-p_s)/(\alpha-1))$$

In each period, there will be a Nash equilibrium according to Bertrand competition with vertical differentiation resulting in prices of $\hat{p}_s$ and $\hat{p}_u$ and per period profits of $\hat{\pi}_s$ and $\hat{\pi}_u$.

Importantly, note that, under competition, the total amount of harm is $(1-\beta)(1-F((\hat{p}_u - \hat{p}_s)/(\alpha-1)))E$ rather than $(\beta(1-F(p_u^*)) + (1-\beta)(1-F(p_u^*/\alpha))E$ when the for-profit entity faces no competition. Thus, the socially-minded entity is able to both sell the safe product to some share of users and also to create conditions under which the for-profit entity may raise prices from π_u to $\hat{p}_u$ for the unrestricted product. While the precise conditions that demarcate this possibility are not derived here, the consequence is that the for-profit entity may not be able to preempt competition from the socially-minded entity ex post.

6 Conclusion

This paper has examined whether a socially-minded entity might be able to control AGI if the technology underlying it is non-excludable. It is shown that either entity will be preempted by a for-profit firm, limiting both its ability and incentive to mitigate unrestricted AGI products that emerge in equilibrium. In other words, it demonstrates that a form of 'self-regulation' is unlikely to be effective.

This leads to a natural conclusion that regulating the harmful effects of AGI, or any other technology for that matter, is likely to be conducted by government policy. Indeed, if such policies can create conditions by which externalities are internalised by for-profit firms (for instance, through product liability laws), it will likely bolster conditions by which socially-

minded activities could prove more effective. That, however, is a subject left to future research. Also left for future research is whether the government intervention required in this case could remain under the control of national governments or whether it requires a global effort. The results in this paper point, for the same reasons as given in this paper, to the latter.[13]

[13]Dum Dum Da ... Ta Dum ... Dum Dum Da ... Ta Dum

References

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Christian, B. (2021). *The alignment problem: How can machines learn human values?* Atlantic Books.

Crifo, P. and Forget, V. D. (2015). The economics of corporate social responsibility: A firm-level perspective survey. *Journal of Economic Surveys*, 29(1):112–130.

Deneckere, R. J. and Preston McAfee, R. (1996). Damaged goods. *Journal of Economics & Management Strategy*, 5(2):149–174.

Francois, P. (2007). Making a difference: Labor donations in the production of public goods. *Rand Journal of Economics*, 37(3):714–732.

Fudenberg, D. and Tirole, J. (1985). Preemption and rent equalization in the adoption of new technology. *The Review of Economic Studies*, 52(3):383–401.

Gans, J. S. (2001). Regulating private infrastructure investment: optimal pricing for access to essential facilities. *Journal of Regulatory Economics*, 20:167–189.

Gans, J. S. (2018). Self-regulating artificial general intelligence. Technical report, National Bureau of Economic Research.

Gilbert, R. J. and Newbery, D. M. (1982). Preemptive patenting and the persistence of monopoly. *The American Economic Review*, pages 514–526.

Hansmann, H. (2000). *The ownership of enterprise*. Harvard University Press.

Ho, K. (2020). The messy, secretive reality behind openai's bid to save the world. *MIT Technology Review*.

Jones, C. I. (2023). The ai dilemma: Growth versus existential risk. Technical report, National Bureau of Economic Research.

Katz, M. L. and Shapiro, C. (1987). R and d rivalry with licensing or imitation. *The American Economic Review*, pages 402–420.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Tirole, J., Moisson, P.-H., and Dubois, P. (2023). The (ir) relevance of the cooperative form.