A tale of two Regulatory Markets: the role of institutional incentives in supporting sustainable Regulatory Markets for future AI systems

Paolo Bova[1,2], Alessandro Di Stefano[1] and The Anh Han[1]

[1]School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, TS1 3BA, UK
[2]paolobova@protonmail.com

Introduction

In the near and long term, the deployment of powerful AI capabilities raises concerns of accidents, misuse, and systemic risk (Brundage et al., 2018; Shevlane and Dafoe, 2019; Zwetsloot and Dafoe, 2019; Hernández-Orallo et al., 2019). These capabilities also require new techniques to audit and certify (Cihon et al., 2021a; Gursoy and Kakadiaris, 2022).

Regulatory Markets could help AI governance to be more adaptive (Clark and Hadfield, 2019). Governments set targets and mandate that companies employ the services of private regulators to demonstrate compliance with those targets. Private regulators must compete with each other to regulate AI companies. This competition may lead to innovations in methods to detect unsafe behaviour and better understand what safe development practises look like.

While the size of these benefits is uncertain, regulators must be incentivised to invest in better methods in the first place. One can ask what role governments can play in providing incentives for higher quality regulators to join the regulatory market. To this end, this extended abstract highlights findings from a recent evolutionary game analysis. The paper explores how different institutional incentives influence the evolutionary dynamics of interactions between AI companies and regulators (Bova et al., 2023). Namely, the paper considers two types of incentives governments might consider, showing that only one of these types, dubbed "Vigilant Incentives", can support regulators in evaluating cutting-edge AI systems.

Model

The model includes two well-mixed populations of fixed size, Z_{ai} and Z_{reg}. AI companies compete to be the first to bring new AI capabilities to market. The model of this competition closely mirrors the DSAIR model from Han et al. (2021) where companies can choose to always develop AI systems safely, **AS**, or to always skip those efforts, **AU**. The model extends the DSAIR model to allow companies to observe the quality of a regulator's efforts. AI companies can therefore use the conditional strategy **VS**: companies only develop AI systems safely if they observe a high level of effort from regulators in vetting AI systems.

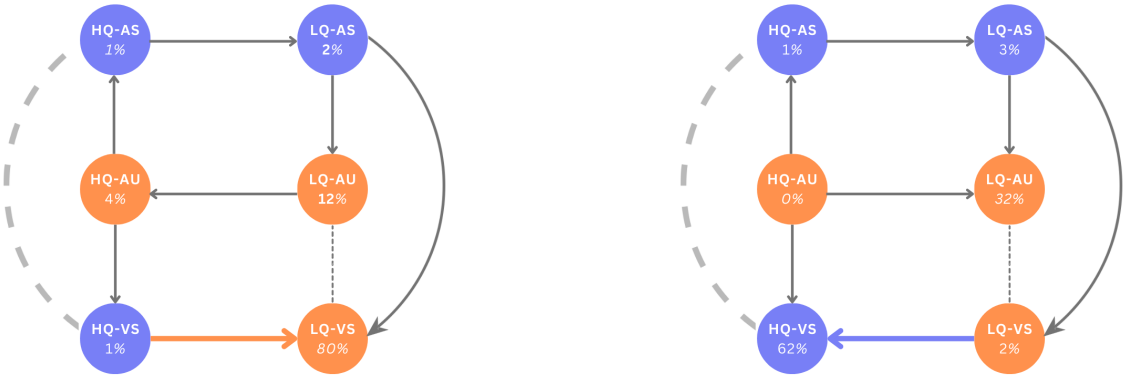
Strategy	Always Safe (AS)	Always Unsafe (AU)
AS	$\frac{B}{2W}$	$p_h(1-\phi)\frac{B}{W}$
AU	$\frac{B}{W}(p(1-p_h)s+p_h\phi)$	$\frac{B}{2W}(p(1-p_h^2)s+p_h^2\phi)$

	Adversarial		Vigilant	
Strategy	HQ	LQ	HQ	LQ
All **AS**	r_h	r_l	r_h+g	r_l+g
All **AU**	r_h+gp_h	r_l+gp_l	$r_h+gp_h{}^2$	$r_l+gp_l{}^2$
All **VS**	r_h	r_l+gp_l	r_h+g	$r_l+gp_l{}^2$

A dilemma may arise where AI companies choose **AU** when society would find this behaviour to be far too reckless. The first-mover advantage when deploying more capable AI systems, B, could be very large, and the timeline until discovery, W, could be relatively short. Thus, even a socially undesirable risk of disaster, p, may not dissuade AI companies from taking shortcuts to reach the discovery at a faster speed, s (Askell et al., 2019; Han et al., 2020). Governments may wish to compel unsafe companies to implement safety measures, slowing them down to a safe speed, ϕ, but they can only do this in time if it is spotted early by a regulator.

Regulators first choose how much to invest in evaluating emerging AI capabilities. A high-quality regulator, **HQ**, accepts higher costs. They have a better chance of detecting unsafe practises in cutting-edge AI systems, $p_h > p_l$, but may earn less overall, $r_h < r_l$. Low-quality regulators, **LQ**, do not invest in detecting new risks. Unable to keep up with new AI capabilities, they are unlikely to detect unsafe behaviour in cutting-edge AI systems (company payoffs in this case are as shown in the table if one replaces p_h with the detection rate for low-quality regulators, p_l). After making their choice, the regulators are randomly matched with two competing AI companies.

Governments want to encourage more Regulatory Market entrants to play **HQ**, but have difficulty assessing their



(a) Adversarial Incentives

(b) Vigilant Incentives

Figure 1: **Only Vigilant Incentives encourage regulators to be high-quality innovators**. On the other hand, Adversarial Incentives allow unsafe AI companies to exploit the presence of low-quality regulators. The Markov Chain diagrams shows the transitions between states and their long-term frequencies. States are coloured blue if AI companies act safely and orange if AI companies act unsafely. The parameters chosen ensure companies act unsafely in the presence of low-quality regulation, although society would prefer them to be safe, $p_h = 0.6$, $p_l = 0$, $g = 1.8$, $\phi = 0.5$, $p_r = 0.6$, $s = 1.5$, $B/W = 100$, $\beta = 0.02$.

quality. They might consider one of two incentives: 'Adversarial Incentives' award g only to regulators who detect unsafe firms. 'Vigilant Incentives' always award g to regulators, unless they later find their matched companies had acted unsafely.

Results and Discussion

The results in Figure 1 concern the transitions between and long-term frequencies of different absorbing states that the two populations converge to under social learning in the limit of rare mutations (Fudenberg and Imhof, 2006; Santos et al., 2016). Figure 1 considers a scenario in which unsafe behaviour is socially irresponsible, but only high-quality regulators can deter unsafe behaviour.

A key finding is that Adversarial Incentives fail to sustain a Regulatory Market. A government may be attracted to Adversarial Incentives because they would only have to pay regulators for detecting unsafe behaviour. However, this intuition misses important dynamics concerning the responses of AI companies.

The Markov chain diagram, Figure 1a, conveys the dynamics at play for Adversarial Incentives. At first, regulators invest heavily in new detection methods, hoping to collect the Adversarial Incentives from catching unsafe AI companies. Soon enough, those regulators will find that AI companies will either move to the conditional strategy **VS** or play **AS**. In either case, the regulator does not have unsafe companies to catch, so they may not have a choice but to choose **LQ** and cancel further investments. At this point, AI companies playing the conditional strategy **VS** waste no time acting unsafely. Meanwhile, those AI companies that play **AS**, learn that due to regulatory complacency, they can get away with developing new AI capabilities unsafely (and must do so to remain competitive). While regulators in state **LQ-AU** may once again invest in new detection methods, these dynamics mean we spend much more time in state **LQ-VS**.

On the other hand, Vigilant Incentives could sustain a Regulatory Market that deters most unsafe behaviour. Figure 1b reveals that it is now much more common to be in the state **HQ-VS**. Most importantly, regulators in the **LQ-VS** state now have a reasonably strong incentive to switch to being **HQ**. Not all regulators are high quality: AI companies may drift to playing **AS**, leading regulators to become complacent, which in turn allows the partial revival of unsafe behaviour. Moreover, regulators can only maintain a high-quality if AI companies play **VS**, and not **AU**. However, AI companies are mainly deterred from unsafe behaviour.

In general, Bova et al. (2023)'s findings paint a clear illustration from which governments should learn. While paying private regulators to only reveal bad behaviour appears at first glance to be an efficient use of government funding, it fails to deter unsafe behaviour, especially when facing AI companies who play conditional strategies. On the other hand, Vigilant incentives ensure that high-quality regulation is incentive-compatible in the face of sophisticated AI companies.

This analysis only scratches the surface on how to design incentives that support a Regulatory Market for AI. Bova et al. (2023) test the robustness of these findings and discuss how to design Vigilant Incentives that balance risk reduction with concerns of overregulation. Finally, governments could consider policies that may complement Regulatory Markets (O'Keefe et al., 2020; Naudé and Dimitri, 2020; Brundage et al., 2020; Cihon et al., 2021b; Truby et al., 2022).

References

Askell, A., Brundage, M., and Hadfield, G. (2019). The Role of Cooperation in Responsible AI Development. *arXiv*.

Bova, P., Di Stefano, A., and Han, T. A. (2023). Both eyes open: Vigilant incentives help regulatory markets improve ai safety. *arXiv preprint arXiv:2303.03174*.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., et al. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *arXiv*.

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., and Baum, S. D. (2021a). AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries. *IEEE Transactions on Technology and Society*, 2(4):200–209.

Cihon, P. J., Schuett, J., and Baum, S. D. (2021b). Corporate governance of artificial intelligence in the public interest. *Inf.*, 12:275.

Clark, J. and Hadfield, G. K. (2019). Regulatory Markets for AI Safety. *arXiv*.

Fudenberg, D. and Imhof, L. A. (2006). Imitation processes with small mutations. *Journal of Economic Theory*, 131(1):251–262.

Gursoy, F. and Kakadiaris, I. A. (2022). System Cards for AI-Based Decision-Making for Public Policy.

Han, T. A., Pereira, L. M., Lenaerts, T., and Santos, F. C. (2021). Mediating artificial intelligence developments through negative and positive incentives. *PLOS ONE*, 16(1).

Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2020). To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. *Journal of Artificial Intelligence Research*, 69:881–921.

Hernández-Orallo, J., Martínez-Plumed, F., Avin, S., and Heigeartaigh, S. O. (2019). Surveying safety-relevant AI characteristics. In *Aaai Workshop on Artificial Intelligence Safety (Safeai 2019)*, pages 1–9. CEUR Workshop Proceedings.

Naudé, W. and Dimitri, N. (2020). The race for an artificial general intelligence: Implications for public policy. *AI & SOCIETY*, 35(2):367–379.

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., et al. (2020). The windfall clause: Distributing the benefits of AI for the common good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 327–331, New York, NY, USA. Association for Computing Machinery.

Santos, F. P., Encarnação, S., Santos, F. C., Portugali, J., and Pacheco, J. M. (2016). An Evolutionary Game Theoretic Approach to Multi-Sector Coordination and Self-Organization. *Entropy*, 18(4):152.

Shevlane, T. and Dafoe, A. (2019). The offense-defense balance of scientific knowledge: Does publishing AI research reduce misuse? *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Truby, J., Brown, R. D., Ibrahim, I. A., and Parellada, O. C. (2022). A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications. *European Journal of Risk Regulation*, 13(2):270–294.

Zwetsloot, R. and Dafoe, A. (2019). Thinking About Risks From AI: Accidents, Misuse and Structure. Retrieved February 2023 from https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.