Inertia and fear of lagging behind drive unsafe technological development in an idealised AI Race experiment

Elias Fernández Domingos[1,2,a,*] and The Anh Han[4]

[1] AI Lab, Vrije Universiteit Brussel, Belgium
[2] MLG, Université Libre de Bruxelles, Belgium
[3] Center for Digital Innovation, Teeside University, UK
[a]ORCID ID: 0000-0002-4717-7984 [b]ORCIRD ID: 0000-0002-3095-7714
[*]Corresponding author. elias.fernandez.domingos@vub.be

Submission type E. Late Breaking Abstract.

Keywords: AI Race · Behavioural Economics · Evolutionary Game Theory.

The success and popularisation of Large Language Models (LLMs) has shown how easily humans adopt disruptive technologies to assist in decision-making for daily tasks. For example, Potter et al. [2] cites several studies demonstrating that LLMs with radical political leanings can, at least in the short term, influence their users' views. This highlights how the involvement of technologies with inherently obscure (black-box) designs in crucial social and (in this case, democratic) decisions may significantly impact the future configuration of our societies.

The safety measures taken in the development of such technologies are essential to mitigate the risks of unexpected and potentially catastrophic effects (e.g., increased misinformation, biases, and inequality). However, excessive regulation of technological development can also be detrimental for beneficial innovation [1]. Therefore, investigating the incentive mechanisms and intrinsic motivations behind human decisions that lead to either unsafe or safe technological development in a theoretical technology race has the potential to inform the design of tailored regulations and policies.

Here, we present the results of a framed behavioural experiment in which participants take on the role of the head of a company. They make binary decisions over several rounds to develop their company's technology either safely or unsafely in order to win a race for technological supremacy against another player (a two-player game).

Throughout an undefined number of rounds (on average, 10), participants accumulate both round payoffs and development steps. In a single round, the only Nash equilibrium and the social optimum occur when both players choose the unsafe option. Additionally, unsafe development advances the player by 1.5 steps in the game, while safe development advances them by only 1 step. However, with each unsafe choice, a player's private risk increases, meaning that the final accumulated private risk is the fraction of unsafe choices made throughout the game, up to a maximum risk threshold. When the game ends, the player with the most steps wins a bonus payoff. Furthermore, only the winning player is liable for private risk, which could result in a final payoff of zero.

We manipulate the maximum private risk at three levels: 10%, 60%, and 90%. Moreover, we base this experiment on the evolutionary game-theoretical model introduced in [1] and hypothesise that most participants at a 90% maximum risk level will choose to develop their technology safely, while at 60% and 10%, most participants will opt for the unsafe option. However, we find significant differences in the total frequency of unsafe decisions only between the 10% and the other two levels, while no difference is found between 60% and 90%. Moreover, we find that participants are willing to signal from round 1 that they will develop the technology unsafely to win the race, regardless of the manipulated risk. We also find that inertia plays a major role in participant's decisions, as well as the interaction the between the distance in the race and whether the opponent has played unsafe in the previous round (see Table 1). Finally, in contrast to the model of [1], participants often choose conditional strategies that begin with unsafe decisions rather than safe ones. Updating the model with this finding yields results that align more closely with our experimental outcomes.

Overall, we find that participants are willing to take high risks to win the race, underscoring the need for external regulation to prevent catastrophic events associated with unsafe technological development.

Table 1. Mixed-effects logistic regression model predicting the probability of choosing `Unsafe`, with a random intercept for `group_id`.

Fixed Effect	**Estimate**	**Std. Error**	**z value**	**Pr($>|z|$)**
Intercept	0.7577	0.1907	3.973	$< 0.001^{***}$
Max. Private Risk	-0.4423	0.2368	-1.868	$0.0618^{\cdot}$
$a_i(t-1)$: Unsafe	-0.6181	0.1484	-4.164	$< 0.001^{***}$
$a_{-i}(t-1)$: Unsafe	0.2659	0.1488	1.787	$0.0740^{\cdot}$
$\Delta S(t-1)$	-0.3083	0.1579	-1.952	$0.0509^{\cdot}$
Unsafe × Unsafe (interaction)	0.2982	0.1995	1.495	0.1350
Unsafe × $\Delta S(t-1)$	0.3982	0.2062	1.931	$0.0535^{\cdot}$
Unsafe_other × $\Delta S(t-1)$	0.4160	0.2104	1.977	0.0480^{*}
3-way Interaction (Unsafe × Unsafe_other × $\Delta S(t-1)$)	-0.1937	0.2619	-0.740	0.4594

Note. Random intercept for group_id: SD = 0.8089.
Significance codes: f$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Acknowledgments. E.F.D. is supported by an F.W.O. Senior Postdoctoral Grant (12A7825N). T.A.H. is supported by EPSRC (grant EP/Y00857X/1). The experiments were funded through the generous support from the Future of Life institute (TAH).

Disclosure of Interests. No competing interests.

References

1. Han, T.A., Pereira, L.M., Santos, F.C., Lenaerts, T., et al.: To regulate or not: a social dynamics analysis of an idealised ai race. Journal of Artificial Intelligence Research **69**, 881–921 (2020)
2. Potter, Y., Rand, D., Choi, Y., Song, D.: Llms' potential influences on our democracy: Challenges and opportunities. In: The Fourth Blogpost Track at ICLR 2025 (2024)