Against Purposeful Artificial Intelligence Failures

Roman V. Yampolskiy
Computer Science and Engineering
University of Louisville
roman.yampolskiy@louisville.edu

"*Our best hope may lie in a succession of small calamities which may teach a more convincing lesson than the preaching of Doomsday prophets.*"
- Adolph Lowe

"Poorly argued claims that superintelligence cannot reasonably pose an existential risk to humanity will make some AI researchers so mad that they will unleash unsafe superintelligence just to make a point."
- Joscha Bach

Abstract
Thousands of researchers are currently of the opinion that advanced artificial intelligence could cause significant damage if developed without appropriate safety measures, but such measures are not currently deployed or even developed. A fringe theory suggests that a severe AI accident could serve as a fire alarm for humanity to take existential dangers of AI seriously and so it is desirable to create such a failure on purpose ASAP to prevent greater harm in the future. In this paper we rely on analogy to inoculation theory to argue against creating purposeful AI failures.

Keywords: *AI Accident, AI Failure, AI Safety, AI Terrorism, Superintelligence, X-Risk.*

1. Introduction
In recent years, the rapid advancements in Artificial Intelligence (AI) have sparked intense debate regarding the future implications of increasingly powerful and autonomous systems [1, 2]. Among the most pressing concerns is the potential for advanced AI to cause significant harm if developed or deployed without appropriate safeguards [3-5]. This anxiety is not unfounded; many researchers and technologists have echoed the sentiment that without rigorous safety measures in place, the advent of superintelligent AI could pose existential risks to humanity [6]. The stakes are further heightened by the current pace of AI development, which often outstrips the formulation and implementation of safety protocols. Despite that, a significant group of AI risk skeptics remains unconvinced of likely dangers [7, 8].

Amid this backdrop of concern and skepticism, a fringe theory has emerged, positing that a purposeful AI failure might serve as a necessary and effective catalyst for action. Proponents of this view argue that a controlled, AI accident could function akin to a "fire alarm," [9] jolting humanity into a state of heightened awareness and urgency regarding the existential dangers of AI. The underlying premise is that only through experiencing the immediate and tangible consequences of an AI failure will global stakeholders be motivated to commit the necessary resources and attention to develop and enforce robust AI safety measures. Such fire alarm would

essentially be a false flag operation [10] in which AI is blamed for human action, in the hopes of inducing anti-AI response, similar how some terrorist acts have been perpetrated to start wars.

In general, preemptive work on malevolent AI [11, 12] may be sufficient to trigger additional regulation and safety research. "Let's build an AI that is power seeking to see what kinds of resources it tries to obtain. Let's build an AI that can simulate human consciousness, emotion, and drive and see how good it can be at persuading people that it is alive. Let's see how good it can be at manipulating people Let's see what kinds of drives it comes up with that might not be aligned with humans (or might consider humans an obstacle to its goals). Let's let it propose changes to its architecture and code and see what kinds of problems that might create. Let's build all of this in a safe, secure environment so we really know what level of risk we are facing, and more importantly, so we build capacity for dealing with these edge risks in a planned, controlled fashion." [13].

It is not uncommon to talk about disasters as opportunities [14], but here is a brief sampling of literature which either implicitly or explicitly suggests that purposeful accidents could be beneficial:

- "A practitioner once suggested that a few "minor" accidents would be desirable to focus the minds of governments and professional organizations on the task of producing safe AI." [15].
- "… only an accident or a near-death experience will jar us awake. … Artificial intelligence is not yet on our existential threat radar. Again, an accident would change that, just as 9/11 introduced the world to the concept that airplanes could be wielded as weapons. That attack revolutionized airline security and spawned a new forty-four-billion-dollar-a-year bureaucracy, the Department of Homeland Security. Must we have an AI disaster to learn a similarly excruciating lesson? Hopefully not, because there's one big problem with AI disasters. They're not like airplane disasters, nuclear disasters, or any other kind of technology disaster with the possible exception of nanotechnology. That's because there's a high probability we won't recover from the first one. … But what kinds of AI-related accidents are we likely to endure on the road to building AGI? And will we be frightened enough by them to consider the quest for AGI in a new, sober light?" [16].
- "Let's call a smaller catastrophe a "warning shot" if it is likely to provoke major useful action to eliminate future risk of that type. Other things being equal, we should prioritize the sharp risks—those that are less likely to be preceded by warning shots—for they are more likely to remain neglected over the long run." [17].
- "If an endurable disaster inspires a response that reduces the chance of even greater future catastrophes or existential risk, then it may be overwhelmingly net-positive in expectation. In this sense, surprising disasters which substantially change our paradigms for risk may not be worth preventing. This should be a major factor considered in the (de)prioritization of certain cause work." [18].

However, this type of suggestion is fraught with ethical, practical, and theoretical challenges. It is also unlikely to work. Drawing on the analogy to inoculation theory [19, 20] —which suggests that exposure to a weakened form of a pathogen can stimulate an immune response without causing full-blown disease—we critically examine the validity and desirability of intentionally creating AI

failures. Inoculation theory, while effective in the context of biological systems, offers a different lesson in the domain of AI: just as a vaccine introduces a controlled risk to build immunity, a purposeful AI failure will introduce a controlled risk with the intention of catalyzing action, but may instead build up tolerance of AI risk and implicit presumption of controllability and survivability of AI failures among the populace.

This paper argues against the proposition of creating purposeful AI failures, highlighting the moral, technical, and practical dilemmas inherent in such an approach. By examining the ethical implications of deliberately inducing harm, the technical challenges of controlling [21] and predicting AI behavior, and the broader impacts on public trust and policy, we advocate for a more responsible path forward.

2. Potential Scenarios for Purposeful AI Accidents

The theoretical consideration of orchestrating a purposeful AI accident to serve as a cautionary signal necessitates an exploration of what such scenarios might entail [11, 13]. These scenarios, while purely hypothetical, underscore the complexity and potential recklessness of attempting to control and predict the outcomes of deliberate AI failures. Understanding the nature of these potential accidents involves considering their scope, the mechanisms by which they could be initiated, and the presumed intentions behind them.

Misaligned AI Systems: One possible scenario for a purposeful accident involves the intentional creation and release of an AI system with known misalignments between its objectives and human values. This could manifest in an AI designed to optimize a certain parameter without adequate safeguards for ethical or safety considerations, leading to unintended and potentially harmful outcomes. The aim would be to demonstrate the critical importance of alignment in AI systems, but the risk is profound: once unleashed, the AI's actions could not be easily predicted or contained, potentially leading to irreversible damage.

Exploitation of AI Vulnerabilities: Another scenario could involve the deliberate exposure and exploitation of vulnerabilities within existing AI systems. This might include, for instance, hacking into autonomous vehicle systems to cause malfunctions, or manipulating recommendation algorithms to spread misinformation at an unprecedented scale. While intended to highlight the fragility and susceptibility of AI systems to malicious interference, such actions would inherently cross ethical boundaries and could have wide-ranging consequences for public safety and trust.

Creation of an AI Provocateur: A more extreme scenario would be the development of an AI with the specific purpose of challenging human control, intended as a stark illustration of the existential risks posed by unbridled AI development. This AI could be tasked with non-destructive but disruptive tasks that underscore its ability to outmaneuver human oversight. The inherent danger in this scenario is the assumption that the AI's capabilities can be precisely calibrated to avoid genuine harm, an assumption that is both hubristic and fraught with uncertainty.

Ethical and Practical Considerations: Each of these scenarios, while diverse in their specifics, shares a common set of ethical and practical challenges. Ethically, the notion of intentionally creating risk, particularly risk that could harm individuals or society, is deeply problematic. It

contravenes the principle of non-maleficence, a foundational element of ethical conduct in both science and technology. Practically, the assumption that a purposeful failure could be tightly controlled is fundamentally flawed. AI systems, especially those involving advanced machine learning and autonomy, can behave in unpredictable ways [22], making it exceedingly difficult to ensure that the consequences of a purposeful accident remain within intended bounds.

In the discourse on the potential use of purposeful AI accidents as a catalyst for broader societal and regulatory engagement with AI safety, a particularly contentious argument emerges: the notion that not just any accident, but a significantly impactful one, is necessary to truly awaken global stakeholders to the urgency of existential risks posed by advanced AI. This line of thought suggests that minor failures or controlled experiments may not suffice to galvanize action; instead, a large-scale, undeniable demonstration of AI's potential for harm might be required.

The argument for a significant, purposeful accident rests on the assumption that the magnitude of an event directly influences its capacity to effect change. It's predicated on the belief that small-scale incidents might be too easily dismissed, rationalized, or absorbed by the existing technological and social systems without necessitating a fundamental reassessment of AI development practices and safety protocols. A large-scale accident, in contrast, could theoretically provide a stark, unignorable demonstration of the need for immediate, comprehensive action to mitigate AI risks.

3. Inoculation Theory and Purposeful Failure

The concept of inoculation, a principle borrowed from immunology, provides a compelling analogy for understanding the debate surrounding purposeful AI failures. In medical science, inoculation involves the introduction of a weakened or dead pathogen to an organism's immune system. This process is designed to provoke a response that prepares the immune system to recognize and combat the pathogen more effectively in future encounters, without causing the full disease. The logic underlying inoculation theory in psychology is similar: exposure to a mild form of a challenge strengthens the individual's or system's ability to handle more significant challenges later on [19, 20]. It's a principle that has found application in areas ranging from health communication strategies to resilience training. At first glance, the idea of applying inoculation theory to AI safety seems to offer an intriguing solution: could a purposeful, controlled failure of an AI system act as a 'vaccine' against more catastrophic failures in the future? Proponents of this theory argue that experiencing a controlled failure could catalyze the global community into action, much like a vaccine prepares the immune system by exposing it to a harmless version of the virus.

Creating an intentional yet relatively minor AI accident or near-miss scenario to try to galvanize AI governance efforts could very easily backfire through an inoculation-style effect. A weak, contained incident may come across as trying to appeal to fear about the risks of advanced AI systems. However, based on inoculation theory, this could perversely reinforce people's resistance to AI regulation rather than raise their perceived threat level to motivate policy action. Specifically, such a stunt could make the public, businesses, and policymakers feel like they've already experienced and weathered an AI incident without catastrophic consequences. Their psychological defenses may kick in to minimize the threat level. They may think to themselves: "Well, an AI

incident happened and it wasn't that bad. The benefits of AI still outweigh the risks, so we don't need restrictive regulations getting in the way of progress." Just like a weak form of a persuasive argument can inoculate people against being persuaded later on, a minor AI incident may provide just enough of a psychological exposure that leads people to double down on complacency surrounding AI safety issues.

Additionally, an intentional AI incident would likely be perceived as alarmist and based on deception, which could significantly undermine trust and credibility surrounding AI governance efforts. It may come across as technologists or activists "crying wolf" and being unduly fatalistic about AI risks. Once this breach of trust occurs, any serious attempts at AI governance frameworks may face an "inoculation effect" of their own, where people are primed to resist what they perceive as anti-AI fearmongering.

There are also pragmatic risks that a supposedly contained AI incident could still lead to real-world negative consequences, potentially causing economic disruptions, damage to property, or even loss of life. Any collateral damage would make the situation much worse in terms of maintaining public trust and initiating good-faith collaborations on AI governance. Ultimately, inoculation theory indicates that a weak, moderately threatening experience does not necessarily raise people's motivation to prevent greater threats - it can have the opposite effect by reinforcing complacency. Given the incredibly high stakes involved with transformative AI systems and the importance of fostering public trust, an intentional AI incident is simply too risky of a gambit to pursue.

Intentionally orchestrating an event that could potentially cause widespread harm, disrupt lives, or even endanger human well-being crosses a critical ethical boundary. It transforms AI safety work from a preventative and protective discipline into one that risks employing the very harm it seeks to prevent. The moral implications of such a strategy are deeply troubling, raising questions about the ends justifying the means and the ethical responsibilities of those involved in AI development and safety research.

Practically, the notion of controlling the scale and impact of a purposeful AI accident is inherently problematic. AI systems, by their nature, can behave in unpredictable and emergent ways, particularly as they grow in complexity and capability. What begins as a controlled, purposeful failure could rapidly escalate beyond intended boundaries, transforming into a genuine, uncontrollable disaster. The belief that one can precisely calibrate the severity of an AI accident overlooks the myriad ways in which AI can interact with and influence complex systems and infrastructures. This unpredictability not only undermines the feasibility of intentionally designing an accident of a specific magnitude but also significantly increases the risk of unintended consequences.

Moreover, transitioning from a purposeful accident to an uncontrollable, natural AI disaster presents a scenario where the initial intentions behind the accident are rendered moot. In such a case, the ethical justification for the initial act—however tenuous—collapses entirely, leaving only the repercussions of a disaster that was intentionally initiated but not effectively contained. This potential for escalation and loss of control underscores the inherent dangers of engaging with the idea of purposeful accidents as a means of influencing AI safety discourse and policy.

In contemplating the prospect of large-scale, purposeful AI accidents, it is crucial to return to the foundational goals of AI safety research: to anticipate, prevent, and mitigate the risks associated with AI development and deployment. Engaging in actions that intentionally introduce risk, particularly on a scale that could lead to widespread harm, is antithetical to these objectives. The pursuit of AI safety must be grounded in ethical principles, a commitment to do no harm, and a proactive approach to identifying and addressing potential risks. Rather than resorting to drastic measures that could themselves constitute an existential threat, the focus should remain on fostering a culture of responsibility, collaboration, and innovation that prioritizes the well-being of humanity in the age of advanced AI.

Individuals and organizations can be influenced to change their behavior in response to significant events or stimuli. Behavioral modeling [23, 24], in this context, refers to the use of theoretical frameworks and empirical data to understand and predict how humans and institutions might respond to various scenarios, including the deliberate introduction of risks or accidents. When considering the concept of purposeful AI accidents as a catalyst for change in AI safety and development practices, behavioral modeling becomes a crucial tool.

4. Empirical Evidence Against the Efficacy of AI Failures as Catalysts

The proposition that a purposeful AI failure could act as a catalyst for global action towards mitigating AI risks hinges on the assumption that such incidents inherently prompt a significant shift in perception and policy regarding AI safety. However, a thorough examination of historical AI failures and their aftermath reveals a more complex and less encouraging picture. The work of Yampolskiy et. al [25-27], provides a comprehensive analysis of numerous AI failures, offering invaluable insights into the actual outcomes of these incidents on public and professional attitudes towards AI development.

Despite a growing catalog of AI mishaps, ranging from relatively benign software glitches to more severe incidents with significant ethical, privacy, and safety implications, the anticipated global mobilization towards rigorous AI safety protocols remains largely unrealized. This observation challenges the core premise of using a purposeful failure as a "fire alarm." The empirical evidence suggests that, rather than catalyzing a concerted effort to address AI risks, past failures have often been met with temporary concern, followed by a rapid return to the status quo. This pattern underscores several critical issues in relying on AI failures to drive policy and perception changes.

First, the nature of technological progress and societal adaptation to new technologies tends to normalize risks and failures. As AI becomes more integrated into daily life and economic activities, there is a tendency for society to become desensitized to the risks, viewing them as an inevitable part of technological advancement. This normalization effect diminishes the impact of individual AI failures to serve as effective wake-up calls. Second, the complexity and abstract nature of AI existential risks make them difficult for the public and policymakers to fully comprehend and internalize. Unlike immediate and tangible dangers, the existential threats posed by superintelligent AI are speculative and contingent on future developments. This makes it challenging to translate the lessons from past AI failures into actionable insights for preventing future existential risks. Third, the fragmentation of the AI research and development landscape complicates the translation of failures into learning opportunities. AI development is a global

endeavor, with numerous stakeholders operating across different jurisdictions, sectors, and ethical frameworks. This diversity, while a source of innovation, also means that lessons from failures are not uniformly disseminated or adopted, diluting the potential for a unified response to AI safety concerns.

Yampolskiy's analysis [25] of AI failures highlights a critical disconnect between the occurrence of these incidents and meaningful advancements in AI safety measures. Rather than acting as catalysts for change, past failures have often led to isolated and superficial responses that fail to address the underlying risks associated with AI development. This empirical evidence strongly argues against the notion that a purposeful failure could effectively galvanize global action towards existential risk mitigation.

5. Conclusions

The exploration of the concept of purposeful AI failures, positioned as a potential catalyst for global action on AI safety, brings to light a multitude of ethical, practical, and theoretical considerations. This paper has sought to dismantle the notion that a controlled, purposeful accident could serve as a beneficial wake-up call, urging humanity to confront the existential risks associated with advanced artificial intelligence. Through the examination of various facets of this proposition—from the analogy with inoculation theory to empirical evidence against the efficacy of past AI accidents as catalysts for change, alongside a discussion on potential purposeful accident scenarios—we arrive at a clear and compelling conclusion: the risks and ethical quandaries inherent in orchestrating a purposeful AI failure far outweigh any speculative benefits.

The analogy with inoculation theory, while initially appealing for its simplicity, is likely to have the undesirable impact of inoculating public against concerns about AI risk. The empirical evidence reviewed further undermines the argument for purposeful AI failures, demonstrating that past accidents have not consistently led to meaningful advancements in AI safety protocols. Rather, these incidents have highlighted the resilience of technological and societal systems to adapt without necessarily addressing the underlying existential threats posed by AI.

Moreover, the discussion of potential scenarios for purposeful AI accidents has illuminated the ethical and practical challenges involved. Each hypothetical scenario underscores the difficulty in predicting AI behavior and the ethical implications of intentionally introducing risk into systems that interact with human lives and societal structures. The potential for unforeseen consequences and collateral damage, both to individuals and to the broader societal trust in AI technologies, presents a formidable argument against such interventions.

In light of these considerations, this paper advocates for a proactive and ethically grounded approach to AI safety. Rather than relying on catastrophic events to prompt action, we must prioritize the development of robust safety measures, ethical guidelines, and governance frameworks that can evolve in tandem with AI technologies. This entails fostering an interdisciplinary dialogue that bridges the gap between technologists, ethicists, policymakers, and the public, ensuring that AI development is guided by a comprehensive understanding of its potential risks and benefits.

The pursuit of AI safety is a collective responsibility that extends beyond the confines of the research community. It requires a concerted effort from all stakeholders involved in the development, deployment, and governance of AI technologies. By embracing a proactive stance on AI safety, informed by ethical principles and empirical evidence, we can navigate the challenges of advanced AI while safeguarding the future of humanity. In this endeavor, the role of purposeful failures, fraught with ethical dilemmas and practical uncertainties, is not only unnecessary but potentially detrimental to the very goals it seeks to achieve.

Acknowledgements

The author is grateful to Jaan Tallinn and the Survival and Flourishing Fund and the Future of Life Institute for partially funding his work. The author would like to thank his assistant, GPT-4, for writing out some of the author's ideas, text summarization, and copyediting.

References

1. Yampolskiy, R.V., *On monitorability of AI.* AI and Ethics, 2024: p. 1-19.
2. Baum, S., A. Barrett, and R.V. Yampolskiy, *Modeling and interpreting expert disagreement about artificial superintelligence.* Informatica, 2017. **41**(7): p. 419-428.
3. Yampolskiy, R.V., *AI: Unexplainable, Unpredictable, Uncontrollable*. 2024: CRC Press.
4. Yampolskiy, R.V., *Untestability of AI. https://www.researchgate.net/publication/378126414_Untestability_of_AI_and_Unfalsifiability_of_AI_Safety_Claims*
5. Brcic, M. and R.V. Yampolskiy, *Impossibility Results in AI: a survey.* ACM Computing Surveys, 2023. **56**(1): p. 1-24.
6. Bengio, Y., et al., *Managing AI risks in an era of rapid progress.* arXiv preprint arXiv:2310.17688, 2023.
7. Yampolskiy, R.V. *AI Risk Skepticism*. in *Conference on Philosophy and Theory of Artificial Intelligence*. 2021. Springer.
8. Ambartsoumean, V.M. and R.V. Yampolskiy, *AI Risk Skepticism, A Comprehensive Survey.* arXiv preprint arXiv:2303.03885, 2023.
9. Yudkowsky, E., *There's No Fire Alarm for Artificial General Intelligence*. October 14, 2017: Available at: https://intelligence.org/2017/10/13/fire-alarm/.
10. Pihelgas, M., *Mitigating risks arising from false-flag and no-flag cyber attacks.* CCD COE, NATO, Tallinn, 2015.
11. Pistono, F. and R.V. Yampolskiy. *Unethical Research: How to Create a Malevolent Artificial Intelligence*. in *25th International Joint Conference on Artificial Intelligence (IJCAI-16). Ethics for Artificial Intelligence Workshop (AI-Ethics-2016)*. 2016.
12. Hubinger, E., et al., *Sleeper agents: Training deceptive llms that persist through safety training.* arXiv preprint arXiv:2401.05566, 2024.
13. Chessen, M., *We Need To Build Doomsday AI*. November 26, 2023: Available at: https://solarpunkfuture.substack.com/p/we-need-to-build-doomsday-ai.
14. Pistono, F., *Coronavirus is a Tragedy. But it May Save Humanity*. March 22, 2020: Available at: https://medium.com/@FedericoPistono/conoravirus-is-tragedy-but-it-may-save-humanity-6f105a1d5fee.

15. Whitby, B., *Reflections on artificial intelligence: the legal, moral and ethical dimensions*. 1996: Intellect Ltd.
16. Barrat, J., *Our final invention: Artificial intelligence and the end of the human era*. 2013: Macmillan.
17. Ord, T., *The precipice: Existential risk and the future of humanity*. 2020: Hachette Books.
18. Anonymous, *The Case for Inspiring Disasters*. June 7, 2020: Available at: https://forum.effectivealtruism.org/posts/zCvWn6f8NdT3mxiZg/the-case-for-inspiring-disasters.
19. McGuire, W.J., *Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments.* The Journal of Abnormal and Social Psychology, 1961. **63**(2): p. 326.
20. Compton, J., *Inoculation theory.* The SAGE handbook of persuasion: Developments in theory and practice, 2013. **2**: p. 220-237.
21. Yampolskiy, R.V., *On the controllability of artificial intelligence: An analysis of limitations.* Journal of Cyber Security and Mobility, 2022. **11**(3): p. 321-404.
22. Yampolskiy, R.V., *Unpredictability of AI: On the impossibility of accurately predicting all actions of a smarter agent.* Journal of Artificial Intelligence and Consciousness, 2020. **7**(01): p. 109-118.
23. Yampolskiy, R.V., *Behavioral modeling: an overview.* American Journal of Applied Sciences, 2008. **5**(5): p. 496-503.
24. Yampolskiy, R.V. and V. Govindaraju. *Use of behavioral biometrics in intrusion detection and online gaming*. in *Biometric Technology for Human Identification III*. 2006. SPIE.
25. Yampolskiy, R.V., *Predicting future AI failures from historic examples.* foresight, 2019. **21**(1): p. 138-152.
26. Scott, P.J. and R.V. Yampolskiy, *Classification schemas for artificial intelligence failures.* Delphi, 2019. **2**: p. 186.
27. Williams, R. and R. Yampolskiy, *Understanding and avoiding ai failures: A practical guide.* Philosophies, 2021. **6**(3): p. 53.