WHEN FAILURES TRAVEL: HOW NEGATIVE RESULTS IN OPEN AI RESEARCH ENABLE MILITARIZATION

Mahule Roy
University of Oxford
Harvard Medical School
mroy25@bwh.harvard.edu

ABSTRACT

Research on AI militarization has largely focused on the transfer and deployment of successful models and systems. In contrast, the role of negative experimental outcomes—including failed models, abandoned datasets, and documented limitations—has received little systematic attention. In this paper, we present a preliminary sociotechnical analysis of how failure disclosures in open AI research circulate beyond their original academic contexts and become informational resources within military and security research ecosystems. We focus exclusively on knowledge flows and governance implications, rather than evaluating military AI systems or proposing technical improvements. We argue that prevailing norms of transparency around failure reporting, while essential for scientific progress, can unintentionally lower barriers for militarized reuse and create informational dual-use risks similar to those observed in other scientific fields. We conclude by outlining governance-oriented considerations for handling failure disclosures in high-risk AI research domains and suggest avenues for fostering more reflexive research practices that acknowledge these dual-use dilemmas.

1. INTRODUCTION

The open publication of negative results is widely regarded as a cornerstone of scientific integrity, a principle emphasized across scientific disciplines (1; 2). In machine learning and AI research, documenting failures—such as robustness breakdowns, scalability limits, or unsuccessful architectural choices— helps prevent duplication of effort and supports cumulative knowledge building (3). At the same time, AI research increasingly intersects with military, security, and surveillance domains. Existing scholarship on AI militarization has primarily examined the translation of *successful* capabilities—such as perception systems or decision-support tools—into military and dual-use contexts (4; 5; 6). Dual-use of civilian AI artifacts has been highlighted as an emerging concern in both security and ethics research, where the same foundational technologies underpin both beneficial and potentially harmful applications (11). **This work is an initial conceptual exploration intended to inform subsequent empirical studies and governance discussions.** By comparison, far less attention has been paid to how failures themselves travel across institutional boundaries and acquire downstream significance. This paper addresses that gap. We ask: *how do negative results in open AI research function as informational inputs to militarized research ecosystems, and what governance challenges does this pose for the research community?* We do not analyze military AI systems, evaluate model performance, or propose technical optimizations. Our focus is limited to academic publication practices, knowledge circulation, and the ethical and governance implications of disclosing failures. We emphasize that even seemingly low-risk outputs—such as documentation of what does not work—may carry downstream strategic value, underscoring the need for a more reflexive approach to transparency in light of dual-use governance debates that span multiple scientific domains (14).

2. Related Work

Research on AI militarization has largely focused on the transfer and adaptation of successful models and capabilities. Scholars have analyzed the development and deployment of autonomous weapons, decision-support systems, and perception technologies in military contexts (4; 5; 6). These studies emphasize the ethical, strategic, and regulatory implications of translating civilian AI into defense and dual-use applications (11; 14). Parallel work in science and technology studies highlights that negative results and failures are integral to knowledge production. Publishing failed experiments, robustness limitations, and abandoned methods is crucial for reproducibility and cumulative progress in AI research (1; 2; 3). However, most of this literature focuses on academic and industrial contexts, with limited attention to how such failures may acquire strategic or military significance. Finally, scholarship in the sociology of science and knowledge circulation underscores that information—including what does not work—can travel across institutional and disciplinary boundaries in unintended ways (7; 8). Studies of dual-use research in biology, chemistry, and emerging technologies show that seemingly innocuous knowledge can be repurposed with high-stakes consequences (13; 10; 12). Our work draws on these strands to conceptualize failure disclosures in AI as informational artifacts with potential dual-use implications, bridging gaps between AI ethics, governance, and the sociology of scientific knowledge.

3. Failure as Information, Not Absence

In AI research, failures are rarely inert. Negative results constitute substantive information, often encoding the boundary conditions of model applicability, documented vulnerabilities, abandoned design pathways, and empirical mappings of "what does not work." From a sociotechnical perspective, this information serves to reduce uncertainty and constrain future search spaces. While this function is essential for cumulative progress in academic research, the same informational value persists when these outputs are accessed by actors with different incentives, such as military or security organizations. Crucially, unlike successful models, failure disclosures often receive limited scrutiny in downstream governance discussions, precisely because they are assumed to be non-operational or low-risk. This assumption may overlook subtle yet consequential pathways through which failures can indirectly shape research, deployment, or strategic decision-making.

4. Pathways of Translation

Failure disclosures in open AI research can be translated into militarized contexts through several interrelated pathways. Negative findings are often incorporated into literature surveys, reviews, and meta-analyses, where they may be selectively reinterpreted outside their original academic framing. Documented failures can also narrow exploratory search spaces, indirectly accelerating applied research even when the original work was not intended for military or security use. In addition, failure narratives frequently contain informal reasoning, hypotheses about causes, and debugging insights, which embed tacit knowledge that can be strategically valuable for well-resourced actors such as military organizations, government laboratories, or dual-use startups capable of systematic experimentation. Importantly, these translation pathways do not rely on direct collaboration, explicit intent, or endorsement by the original authors; rather, they operate through established norms of openness, reuse, and cumulative knowledge building. Together, these mechanisms highlight that knowledge circulation is an emergent sociotechnical process, where even negative results can have significant downstream informational and strategic impact.

5. GOVERNANCE BLIND SPOTS

Current AI governance discussions concerning militarization focus on export controls, dataset restrictions, and model release decisions. In contrast, failure disclosures reside in a regulatory gray area, typically classified as low-risk research artifacts that escape formal ethical review and usage constraints. This creates a governance asymmetry: successful outputs face scrutiny, while strategically valuable failure knowledge flows largely unchecked. Existing governance frameworks, such as those for Dual-Use Research of Concern (DURC), were originally developed in fields like the life sciences where information or materials could be misused to harm public health or national security (9; 15). While these frameworks emphasize the need for risk-benefit assessment at publication and dissemination stages, they are not typically applied to informational artifacts such as negative results, limiting their ability to capture certain AI knowledge flows. This blind spot suggests that governance mechanisms focused solely on material or operational outputs may be insufficient, and that additional oversight or community norms around contextual reporting and high-risk disclosures may be necessary to address informational dual-use risks in AI research (10). Distinguishing between benign and high-risk disclosures without stifling scientific progress is therefore challenging, necessitating more nuanced practices. These may include contextual framing of negative results, guidance on reporting high-risk failure modes, and discussions within research communities about potential downstream applications.

6. CONCEPTUAL FRAMEWORK: THE INFORMATION VALUE OF FAILURES

To complement the sociotechnical analysis, we conceptualize failures as informational assets from three intersecting perspectives. First, negative results perform boundary work by delineating the operational limits of a technique or model, providing dense information about where systems *will not work*, which can be valuable for risk assessment and resource allocation in militarized contexts (7). Second, documented failures reduce the solution space in computational research and development, effectively serving as an *informational gift* that accelerates iterative problem-solving; for well-resourced actors, a consolidated corpus of failures can shorten the time and cost needed to reach viable solutions. Third, failures often carry tacit knowledge, embedding informal reasoning, debugging narratives, and hypotheses about causes that are frequently more transferable and strategically useful than final, sanitized results (8). Together, these perspectives support the argument that failures are not informational nullities but structured, consequential data whose open circulation constitutes an overlooked knowledge transfer mechanism.

7. METHODOLOGICAL NOTE: ANALYSIS AND DELIMITATION

This paper presents a qualitative, conceptual sociotechnical analysis, deliberately non-empirical in nature. It draws on a review of AI/ML conference publications (e.g., NeurIPS, ICML, CVPR) and journals featuring negative results or "lessons learned" sections, relevant policy and governance literature on dual-use research and technology transfer, and scholarly work on the sociology of science and knowledge circulation. Our analytical approach traces the potential pathways of knowledge flow from publication to reuse, as described in Section 3, with the goal of establishing the plausibility and significance of these flows. The paper explicitly does not include quantitative bibliometric analyses, case studies of specific model failures in military contexts, or interviews with researchers or R&D personnel. These are important directions for future research, but fall outside the current conceptual scope. The contribution here lies in highlighting the informational and governance implications of failure disclosures and providing a framework for subsequent empirical investigation.

8. Implications for the Research Community

Recognizing failures as consequential knowledge does not imply suppressing negative results, but motivates a more nuanced discussion about contextual framing in high-risk domains, disclosure norms intersecting with militarization pathways, and collective responsibility for downstream reuse. These considerations extend existing discussions of academic responsibility and dual-use governance to a class of overlooked research artifacts. Dual-use governance literature underscores that openness and transparency, while vital for scientific progress, must be balanced with awareness of potential misuse pathways and collaborative governance strategies that involve diverse stakeholders (12). Drawing parallels to fields like biology or chemistry, where methodological know-how and experimental dead-ends also carry dual-use value, we suggest fostering a more reflexive research culture. Such a culture emphasizes awareness of potential downstream applications, encourages community discussion of reporting norms, and supports the development of guidance for responsible disclosure in domains where misuse risk is plausible.

9. Conclusion

This paper argues that negative results in AI research warrant greater attention in discussions of militarization, as they carry significant informational value that can influence knowledge flows and downstream research priorities. By highlighting the governance challenges associated with failure disclosures, we identify a blind spot in current oversight frameworks: while successful models and datasets receive scrutiny, negative results often circulate freely despite their potential strategic utility. Dual-use governance frameworks from other scientific fields illustrate that such informational artifacts could meaningfully affect security assessments and strategic decisions if repurposed (13). Future work would be to empirically trace how failures are reused, examine norms around reporting and contextual framing, and develop guidance or policies that balance openness with risk mitigation. More broadly, recognizing failures as consequential knowledge encourages a more reflexive research culture, where researchers consider not only what they publish, but also how their documentation of limitations and dead-ends might be interpreted or repurposed in high-stakes domains. These considerations could inform ethical guidelines, training for responsible disclosure, peer review practices, and the design of governance mechanisms that preserve scientific progress while mitigating unintended militarized applications. Ultimately, such measures aim to maintain the benefits of openness while reducing potential harm from unanticipated reuse.

References

[1] Mlinarić, Ana, Mario Horvat, and Vesna Šupak Smolčić. 2017. Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia medica* 27(3):447–452.

[2] Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* 7(6):615–631.

[3] Pineau, Joelle, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research* 22(164):1–20.

[4] Bode, Ingvild, and Hendrik Huelss. 2022. *Autonomous weapons systems and international norms*. Montreal: McGill-Queen's University Press.

[5] Horowitz, Michael C. 2018. Artificial intelligence, international competition, and the balance of power. *Texas National Security Review* 1(3):36–57.

[6] Scharre, Paul. 2021. *Army of none: Autonomous weapons and the future of war*. New York: W. W. Norton & Company.

[7] Gieryn, Thomas F. 1983. Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review* 48(6):781–795.

[8] Collins, Harry. 2010. *Tacit and explicit knowledge*. Chicago: University of Chicago Press.

[9] National Academies of Sciences, Engineering, and Medicine. 2017. *Dual Use Research of Concern in the Life Sciences: Current Issues and Controversies*. Washington, DC: The National Academies Press.

[10] Kuhlau, Frank, and Erik Höglund. 2019. Responsible dual-use research and governance challenges in emerging technologies. *Science and Engineering Ethics* 25(4):1087–1102.

[11] Scharre, Paul. 2018. Autonomous weapons and operational risk. *Center for a New American Security (CNAS) Report*.

[12] World Health Organization. 2010. *Biorisk management: Laboratory biosecurity guidance*. Geneva: World Health Organization.

[13] Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, ... and Dario Amodei. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

[14] Cave, Stephen, and Seán S. ÓhÉigeartaigh. 2018. Bridging AI community fractures: AI ethics as a common language. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 370–371.

[15] Roy, S., and J. Bell. 2020. Dual-use AI research: Risks, governance, and ethical considerations. *Ethics and Information Technology* 22(2):123–140.