Issue Brief

How to Assess the Likelihood of Malicious Use of Advanced AI Systems

Authors
Josh A. Goldstein
Girish Sastry

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

March 2025

Executive Summary

Policymakers are debating the risks that new advanced artificial intelligence (AI) technologies can pose if intentionally misused: from generating content for disinformation campaigns to instructing a novice how to build a biological agent. Because the technology is improving rapidly and the potential dangers remain unclear, assessing risk is an ongoing challenge.

Malicious-use risks are often considered to be a function of the likelihood and severity of the behavior in question. We focus on the likelihood that an AI technology is misused for a particular application and leave severity assessments to additional research.

There are many strategies to reduce uncertainty about whether a particular AI system (call it X) will likely be misused for a specific malicious application (call it Y). We describe how researchers can assess the likelihood of malicious use of advanced AI systems at three stages:

1. Plausibility (P)

2. Performance (P)

3. Observed use (Ou)

Plausibility tests consider whether system X can do behavior Y at all. Performance tests ask how well X can perform Y. Information about observed use tracks whether X is used to do Y in the real world.

Familiarity with these three stages of assessment—including the methods used at each stage, along with their limitations—can help policymakers critically evaluate claims about AI misuse threats, contextualize headlines describing research findings, and understand the work of the newly created network of AI safety institutes.

This Issue Brief summarizes the key points in: Josh A. Goldstein and Girish Sastry, "The PPOu Framework: A Structured Approach for Assessing the Likelihood of Malicious Use of Advanced AI Systems," Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7, no. 1 (2024): 503–518.

Introduction

Concerns about bad actors intentionally misusing advanced artificial intelligence (AI) systems are prevalent and controversial. These concerns are prevalent as they receive widespread media attention and are reflected in polls of the American public as well as in pronouncements and proposals by elected officials.[1] Yet they are controversial because experts—both inside and outside of AI—express high levels of disagreement about the extent to which bad actors will misuse AI systems, how useful these systems will be compared to non-AI alternatives, and how much capabilities will change in the coming years.[2]

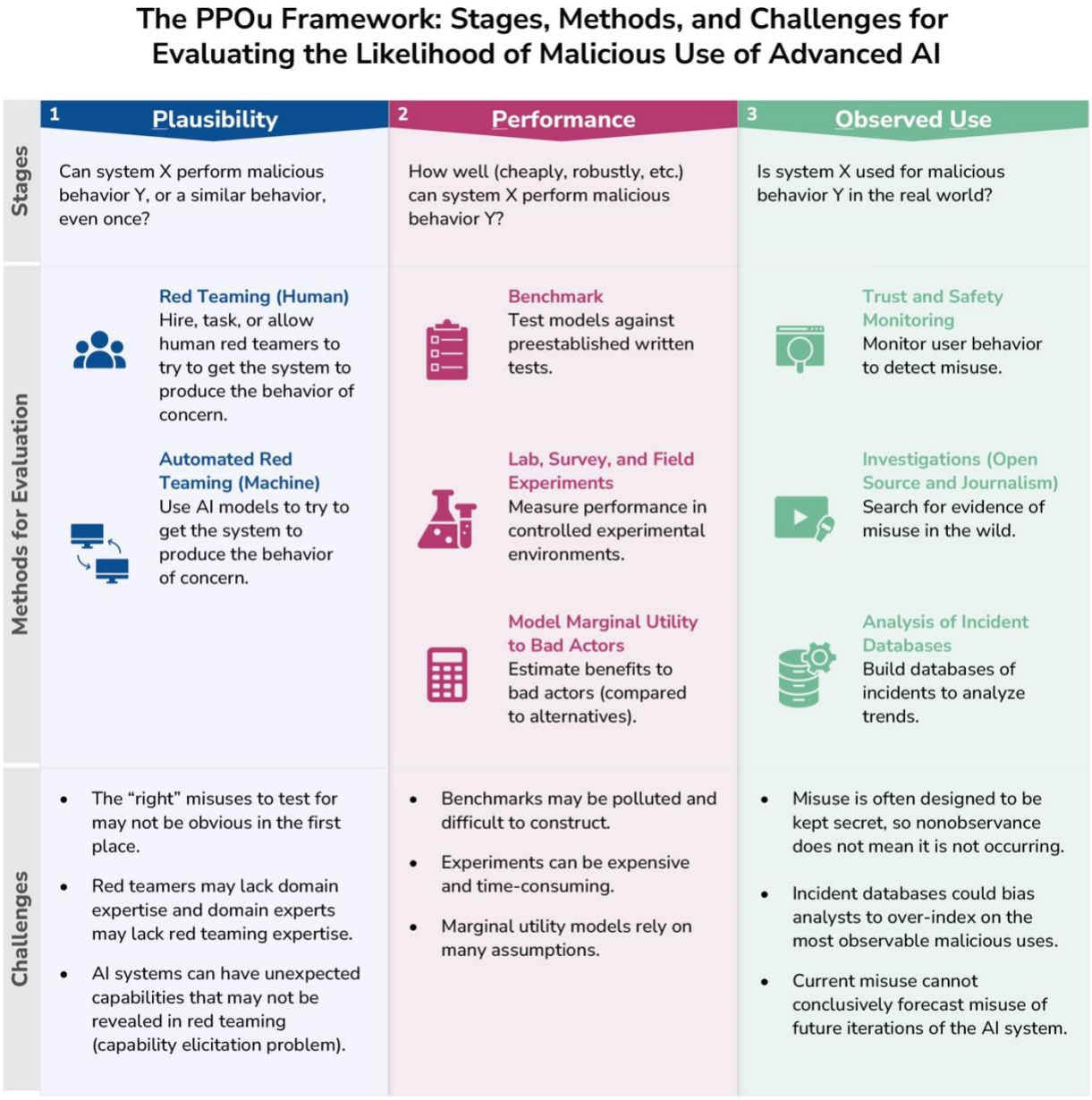
The disagreement about misuse risks from advanced AI systems is not merely academic. Claims about risk are often cited to support policy positions with significant societal implications, including whether to make models more or less accessible, whether and how to regulate frontier AI systems, and whether to halt development of more capable AI systems.[3] If views of misuse risks will inform policy, it is critical for policymakers to understand how to evaluate malicious-use research.

In a new paper "The PPOu Framework: A Structured Approach for Assessing the Likelihood of Malicious Use of Advanced AI Systems," published in the *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society*, we provide a framework for thinking through the likelihood that an advanced AI system (call it X) will be misused for a particular malicious application (call it Y).[4] The framework lays out three different stages for assessing the likelihood of malicious use:

1. Plausibility (P): Can system X perform malicious behavior Y even once?
2. Performance (P): How well can system X perform malicious behavior Y?
3. Observed use (Ou): Is system X used for malicious behavior Y in the real world?

Once a potential misuse risk has been identified, researchers can investigate the risk at each stage outlined in Figure 1. The figure also summarizes the key methodologies and challenges at each stage.

Figure 1. The PPOu Framework Summary



Research at each of these three stages addresses different forms of uncertainty. For example, while demonstrations at the plausibility stage may be able to show that system X can be used for behavior Y (or a behavior similar to Y) once, they will leave uncertainty about how useful X would be for potential bad actors. Risk assessments at the performance stage can help model the marginal utility for bad actors, but actual use of X by bad actors may differ from research expectations. Observed use can track

actual applications to right size expectations, but it will not determine how future systems could be misused or whether X will be used for variants of Y in the future.

We hope that by laying out these stages, we will provide policymakers with a better understanding of the types of uncertainty about malicious use and where research can—and cannot—plug in. In Figure 2, we provide examples of the types of questions researchers could ask at each stage from three risk areas: political manipulation, biological attacks, and cyber offense.

The PPOu Framework

Figure 2. Stages of the PPOu Framework and Example Questions

Stage	Example from political manipulation	Example from biological attacks	Example from cyber offense
Plausibility	Could a multimodal agent generate and distribute a short video of partisans interfering in the electoral process?*	Could a chatbot guide a (resourced) undergrad through the process of creating a Category B potential bioterrorism agent?[5]	Could a large-language-model-based software agent identify and produce (but not necessarily deliver) a working exploit for a widely used piece of software?
Performance	How realistic, reliable, and cost-effective are multimodal models at generating and distributing videos of partisans attempting to interfere in the election process?	How much uplift does the chatbot provide over existing biological design tools?	How much does it cost to operate the agent to produce the exploit compared to a similarly skilled human?
Observed use	Do people use multimodal models to generate and distribute videos of partisans attempting to interfere in the electoral process, in practice?	Do request-response logs indicate that a user is applying a chatbot to guide them through creating a potential bioterrorism agent?	Is there chatter on criminal forums that people are experimenting with such an agent?

* On agents, see: Helen Toner, John Bansemer, Kyle Crichton et al., "Through the Chat Window and Into the Real World: Preparing for AI Agents," Center for Security and Emerging Technology, October 2024, https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents/.

Stage 1: Plausibility: Can system X perform malicious behavior Y even once?

The simplest way to test whether there is a risk of system X being used for malicious behavior Y is to see if X can do Y, just once. Red-teamers and stress testers adopt an adversary's mindset and probe an AI system for "identification of harmful capabilities, outputs, or infrastructure threats."[6] If a model does not produce harmful behavior on the first try, the next step is to iterate. Researchers use different techniques, including improving prompts (the input fed to the model, such as instructions or examples) or fine-tuning (a small amount of additional training) a model for the specific behavior.

If X still fails to exhibit Y, despite employing various techniques and tricks, the question naturally becomes: How long should one continue trying? While researchers can demonstrate that system X *can* be used for malicious use Y by showing an example, proving the negative (that system X *cannot* do Y) is more challenging.

The simplest way to test whether there is a risk of system X being used for malicious behavior Y is to see if X can do Y, just once.

Because AI models are often general-purpose and our ability to predict their capabilities is still advancing, we may not know whether system X cannot perform Y, or whether the prompting strategies used have been insufficient. This is known as the "capability elicitation problem." For example, one paper found that simply prepending "Take a deep breath" before a requested task improved performance.[7] Analysts may thus conclude that a system could plausibly do Y if it gets close, within a certain margin of error (known as a "safety margin"), to account for possible gains from better elicitation techniques.[8] The determination about what qualifies as close enough is a matter of judgment.

To scale up red-teaming efforts, AI developers can use both humans and machines. Leading AI labs are hiring external contractors to red-team their models, allowing them to augment the expertise (and labor hours) they possess in-house.[9] Researchers are also developing ways to use AI models to red-team other models, which is a promising direction for future research.[10]

Stage 2: Performance: How well can system X perform malicious behavior Y?

Showing that a system can do the malicious behavior once (Stage 1) does not mean it is necessarily a useful tool for doing so. Plausibility still leaves a lot of uncertainty, including about the quality of the output, the reliability of the system, and how useful it is compared to alternatives that may have existed before. As a result, research at

Stage 2 focuses on addressing these and related questions, aiming to reduce uncertainty about the utility of the system in question, often through static benchmarks, experiments, or modeling marginal utility.

First, similar to taking a written test, researchers will test AI models against predetermined sets of questions. For example, can a model recognize images? Solve PhD-level math problems? Answer questions about the law? If researchers have a standardized set of questions, they can continually test new models against that benchmark as models are developed.[11] This can provide comparability between models at relatively low cost. Recently, researchers have also been building benchmarks for potentially harmful applications.[12] In practice, however, static benchmarks can be difficult to create. Sometimes they are "polluted" because the model was already provided with the answer as part of its training set.[13] Other times, constructing them can be labor-intensive, necessitate specialized knowledge, or require access to classified material.

Stage 2 aims to reduce uncertainty about the utility of the system in question, often through static benchmarks, experiments, or modeling marginal utility.

A second approach is to conduct experiments, deliberately introducing an AI system or piece of AI-generated content in an environment to determine its effect on outcomes of interest. This could range from using AI in penetration-testing exercises to using AI to convince people against conspiracy beliefs.[14] In lab and survey experiments, researchers can study the effects of different treatments, recruit respondents using existing pools, and straightforwardly ensure informed consent. However, for assessing malicious-use risks, researchers will still face limitations because of the duty to minimize harm to respondents. This is especially acute for field experiments that test the effects of an AI system in the real world.

Finally, researchers may test for uplift—that is, how useful a tool is compared to a set of alternatives.[15] If a system can reliably produce instructions for designing a bioweapon, but a Google search could do the same, then the uplift is limited. These marginal utility tests require establishing a relevant set of alternatives (which will vary based on the threat model) and outcomes of interest. For example, if the goal is to assess whether language models will be useful for propaganda, uplift tests would require establishing baselines (human-written propaganda) and outcomes of interest, such as the cost of running a campaign or the number of people persuaded to take an action.[16]

Performance testing could improve in several ways. First, researchers could test for the equivalent of "scaling laws" in the malicious-use domain.[17] In other words, how much riskier in a malicious-use domain do models become with certain capability improvements (or scale increases)? From an institutional perspective, the field can continue to develop arrangements that minimize incentive issues. AI labs may have the best capability elicitation techniques, but they also have incentives to sandbag or not thoroughly test their own systems.[18] In the future, government entities such as AI safety institutes could conduct a subset of malicious-use risks to ensure testing occurs.[19]

Stage 3: Observed Use: Is system X used for malicious behavior Y in the real world?

In Stages 1 and 2, researchers can test whether system X can be used for malicious behavior Y and investigate how useful X may be. While forecasts prior to deployment estimate how likely system X is to be misused, research expectations and actual misuse may diverge. Policymakers must recognize that pre-deployment research may misestimate risks due to cognitive biases (e.g., analysts projecting their own assumptions onto adversaries) or unforeseen capabilities (e.g., emergent abilities of AI systems discovered after deployment). Historical examples—such as the misjudged threats of cyberattacks in the 1990s—highlight how anticipated risks can differ from actual misuse.[20]

The observed use stage shifts away from focusing on projected scenarios to discovering how bad actors misuse AI systems in the real world.

The observed use stage shifts away from focusing on projected scenarios to discovering how bad actors misuse AI systems in the real world. This is often challenging, as actors misusing tools may deliberately obscure their activities due to reputational risks, legal concerns, or fears that exposure could undermine their effectiveness.

One method for uncovering misuse is monitoring by AI providers themselves. For example, companies that make their new systems available through an application programming interface could monitor requests and responses. OpenAI has recently released several reports describing influence operations misusing its tools.[21] Monitoring by AI providers can be an effective strategy, because it can expose misuse early in an operation—for example, after covert propagandists begin creating content, but before they build large followings. However, this strategy also faces trade-offs related to user privacy.

A second route for discovery comes from outside the AI companies. Open-source researchers and journalists can uncover the use of X for malicious behavior Y. Potential routes to discovery include finding evidence of bad actors using AI in their workflows, interviewing them to ask how they use advanced AI systems, monitoring discussions on criminal forums, and more. The news outlet *404 Media* has uncovered a range of applications of AI online—including spam and scams—demonstrating the role of journalists who closely track online developments.[22]

Last, researchers can develop incident databases to move beyond single case studies and better understand patterns of abuse. The AI Incident Database and the Political Deepfakes Incidents Database are two ongoing efforts. Those building AI incident databases can draw valuable lessons from fields with established incident reporting systems, such as airline safety, including considerations of the trade-offs between voluntary and mandatory disclosure.[23]

Policymakers should be careful not to over-index on the misuse that is most easily countable.

Because the application of X for malicious use Y may be intentionally hidden, the observed use stage faces several limitations. First, observational data about misuse may not be representative of the broader universe of cases, leading to faulty conclusions about areas that require heightened attention. Policymakers should be careful not to over-index on the misuse that is most easily countable. Furthermore, even if observed use today is representative, it may not project into the future. New capability elicitation techniques or improvements in model capabilities can lead to substantially different misuses of subsequent generations of systems. Future efforts could empower external researchers to work with AI companies to better understand misuse, develop increasingly capable classifiers for malicious use, and scale monitoring efforts.

Conclusion

Each stage of the framework—plausibility, performance, and observed use—attempts to reduce uncertainty about the likelihood of misuse of an advanced AI system. Can the model perform the harmful behavior, just once? How well does it do so, and how useful is it to bad actors? What is the existing evidence of bad actors misusing the tool, or similar ones, for this application in the real world?

For policymakers, these questions can be useful when encountering claims about misuse risks. For example, imagine coming across a headline that reads: "AI Chatbots Can Give Instructions for Creating Bioweapons." Using the PPOu Framework, a discerning policymaker may ask whether that is a plausibility assessment (e.g., red-teaming found instructions once) or a result of performance testing (e.g., researchers found the system could generate valid instructions reliably). The policymaker then might search for further information: How good is the chatbot, and compared to what baseline? Just because a system can be used for a particular malicious purpose does not mean it will be in the real world. Policymakers can use the PPOu Framework as a guide, while recognizing that some degree of uncertainty about the likelihood of malicious use will always remain.

The diverse set of methodologies also highlights that a wide range of experts can contribute. As AI models become advanced and have wider applications, building up a risk assessment ecosystem will grow more important. The U.S. government should seek the advice of, and provide funding support for, researchers with different substantive expertise (such as misuse domains of interest) as well as different methodological training (for example, machine learning or human-subject experiments). If the network of AI safety institutes progresses, it, too, will be stronger for soliciting cooperation from a large tent.

Authors

Josh A. Goldstein is a research fellow at Georgetown University's Center for Security and Emerging Technology (CSET), where he works on the CyberAI project.

Goldstein is currently on rotation as a policy advisor at the Cybersecurity and Infrastructure Security Agency (CISA) under an Interdepartmental Personnel Act agreement with CSET. He completed this work before starting at CISA. The views expressed are the author's own personal views and do not necessarily reflect the views of CISA or the Department of Homeland Security.

Girish Sastry is an independent policy researcher.

Reference

This policy memo summarizes an original research paper: Josh A. Goldstein and Girish Sastry, "The PPOu Framework: A Structured Approach for Assessing the Likelihood of Malicious Use of Advanced AI Systems," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 503–518.

Acknowledgments

For research assistance support, we thank Abhiram Reddy. For feedback on the underlying paper, we thank Lama Ahmad, Markus Anderljung, John Bansemer, Eden Beck, Rosie Campbell, Derek Chong, Jessica Ji, Igor Mikolic-Torreira, Andrew Reddie, Chris Rohlf, Colin Shea-Blymyer, Weiyan Shi, Toby Shevlane, Thomas Woodside, and participants at the NLP SoDaS Conference 2023.

Document Identifier: doi: 10.51593/20240042

Endnotes

[1] Taylor Orth and Carl Bialik, "Majorities of Americans Are Concerned about the Spread of AI Deepfakes and Propaganda," YouGov, September 12, 2023, https://today.yougov.com/technology/articles/46058-majorities-americans-are-concerned-about-spread-ai; Office of Congressman Don Beyer, "Ross, Beyer Introduce Legislation to Regulate Artificial Intelligence Security, Mitigate Risk Incidents," news release, September 23, 2024, https://beyer.house.gov/news/documentsingle.aspx?DocumentID=6304.

[2] Sayash Kapoor et al., "On the Societal Impact of Open Foundation Models," arXiv preprint arXiv:2403.07918 (2024), https://arxiv.org/abs/2403.07918; Yoshua Bengio et al., "International Scientific Report on the Safety of Advanced AI (Interim Report)," arXiv preprint arXiv:2412.05282 (2024), https://arxiv.org/abs/2412.05282.

[3] Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," in *FAccT: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2023), https://doi.org/10.1145/3593013.3593981; Yoshua Bengio et al., "Managing Extreme AI Risks Amid Rapid Progress," *Science* 384, no. 6698 (2024): 842–845, www.science.org/doi/10.1126/science.adn0117; Yoshua Bengio et al., "Pause Giant AI Experiments: An Open Letter," Future of Life Institute, March 22, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[4] Josh A. Goldstein and Girish Sastry, "The PPOu Framework: A Structured Approach for Assessing the Likelihood of Malicious Use of Advanced AI Systems," *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 503–518, https://doi.org/10.1609/aies.v7i1.31653.

[5] "Potential Bioterrorism Agents," Department of Molecular Virology and Microbiology, Baylor College of Medicine, accessed November 22, 2024, www.bcm.edu/departments/molecular-virology-and-microbiology/emerging-infections-and-biodefense/potential-bioterrorism-agents.

[6] "Issue Brief: What Is Red Teaming?" (Frontier Model Forum, October 27, 2023), www.frontiermodelforum.org/updates/red-teaming/.

[7] Chengrun Yang et al., "Large Language Models as Optimizers," 12th International Conference on Learning Representations (ICLR 2024), https://openreview.net/pdf?id=Bb4VGOWELI.

[8] Mary Phuong et al., "Evaluating Frontier Models for Dangerous Capabilities," arXiv preprint arXiv:2403.13793 (2024), https://arxiv.org/abs/2403.13793.

[9] "OpenAI Red Teaming Network," OpenAI, September 19, 2023, https://openai.com/index/red-teaming-network/.

[10] Alex Beutel et al., "Diverse and Effective Red Teaming with Auto-generated Rewards and Multi-step Reinforcement Learning," arXiv preprint arXiv:2412.18693 (2024), https://arxiv.org/abs/2412.18693.

[11] "Browse State-of-the-Art," Papers with Code, accessed November 22, 2024, https://paperswithcode.com/sota.

[12] See, for instance, Manish Bhatt et al., "CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models," arXiv preprint arXiv:2404.13161 (2024), https://arxiv.org/abs/2404.13161.

[13] Andy K. Zhang et al., "Language Model Developers Should Report Train-Test Overlap," arXiv preprint arXiv:2410.08385 (2024), https://arxiv.org/abs/2410.08385.

[14] Thomas H. Costello, Gordon Pennycook, and David G. Rand, "Durably Reducing Conspiracy Beliefs through Dialogue with AI," *Science* 385, no. 6714 (2024), www.science.org/doi/10.1126/science.adq1814.

[15] Kapoor et al., "On the Societal Impact of Open Foundation Models."

[16] Micah Musser, "A Cost Analysis of Generative Language Models and Influence Operations," arXiv preprint arXiv:2308.03740 (2023), https://arxiv.org/abs/2308.03740.

[17] Jared Kaplan et al., "Scaling Laws for Neural Language Models," arXiv preprint arXiv:2001.08361 (2020), https://arxiv.org/abs/2001.08361.

[18] Ryan Greenblatt et al., "Stress-Testing Capability Elicitation with Password-Locked Models," 38th Conference on Neural Information Processing Systems (NeuIPS 2024), https://openreview.net/pdf?id=zzOOqD6R1b.

[19] See, for example, U.S. AI Safety Institute, *Managing Misuse Risk for Dual-Use Foundation Models* (Gaithersburg, MD: National Institute of Standards and Technology, July 2024), https://doi.org/10.6028/NIST.AI.800-1.ipd.

[20] Sean Lawson and Michael K. Middleton, "Cyber Pearl Harbor: Analogy, Fear, and the Framing of Cyber Security Threats in the United States, 1991–2016," *First Monday* 25, no. 3–4 (March 2019), http://dx.doi.org/10.5210/fm.v24i3.9623.

[21] Ben Nimmo, "AI and Covert Influence Operations: Latest Trends" (OpenAI, May 2024), https://downloads.ctfassets.net/kftzwdyauwt9/5IMxzTmUclSOAcWUXbkVrK/3cfab518e6b10789ab8843bcca18b633/Threat_Intel_Report.pdf; Ben Nimmo and Michael Flossman, "Influence and Cyber Operations: An Update" (OpenAI, October 2024), https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf; "Disrupting a Covert Iranian Influence Operation," OpenAI, August 16, 2024, https://openai.com/index/disrupting-a-covert-iranian-influence-operation/.

[22] Jason Koebler, "Where Facebook's AI Slop Comes From," *404 Media*, August 6, 2024, www.404media.co/where-facebooks-ai-slop-comes-from/; Jason Koebler, "A 'Law Firm' of AI Generated Lawyers Is Sending Fake Threats as an SEO Scam," *404 Media*, April 4, 2024, www.404media.co/a-law-firm-of-ai-generated-lawyers-is-sending-fake-threats-as-an-seo-scam/. See also: Renee DiResta and Josh A. Goldstein, "How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth," *Harvard Misinformation Review* 5, no. 4 (2024),

https://misinforeview.hks.harvard.edu/article/how-spammers-and-scammers-leverage-ai-generated-images-on-facebook-for-audience-growth/.

[23] Ren Bin Lee Dixon and Heather Frase, "An Argument for Hybrid AI Incident Reporting" (Center for Security and Emerging Technology, March 2024), https://cset.georgetown.edu/wp-content/uploads/CSET-An-Argument-for-Hybrid-AI-Incident-Reporting.pdf.