The PPOu Framework: A Structured Approach for Assessing the Likelihood of Malicious Use of Advanced AI Systems

Josh A. Goldstein[*1], **Girish Sastry**[*2]
[1]Georgetown University, Center for Security and Emerging Technology
[2]OpenAI
jg2218@georgetown.edu, girish@openai.com

Abstract

The diffusion of increasingly capable AI systems has produced concern that bad actors could intentionally misuse current or future AI systems for harm. Governments have begun to create new entities—such as AI Safety Institutes—tasked with assessing these risks. However, approaches for risk assessment are currently fragmented and would benefit from broader disciplinary expertise. As it stands, it is often unclear whether concerns about malicious use misestimate the likelihood and severity of the risks. This article advances a conceptual framework to review and structure investigation into the likelihood of an AI system (X) being applied to a malicious use (Y). We introduce a three-stage framework of (1) Plausibility (can X be used to do Y at all?), (2) Performance (how well does X do Y?), and (3) Observed use (do actors use X to do Y in practice?). At each stage, we outline key research questions, methodologies, benefits and limitations, and the types of uncertainty addressed. We also offer ideas for directions to improve risk assessment moving forward.

Introduction

Concern about malicious use of advanced machine learning systems (AMLSs) dates back decades. Recent improvements in natural language processing, however, have brought concern to the fore. Researchers have suggested that machine learning (ML) systems can produce language to generate synthetic term papers (Franke and Alexander 2019), fake news (Zellers et al. 2019), extremist recruiting materials (McGuffie and Newhouse 2020), and more (Bender et al. 2021). Others have outlined threats of AI in digital security (e.g., cyberattacks), physical security (e.g., autonomous weapons), and political security (e.g., persuasion and deception) (Brundage et al. 2018). With the widespread rollout of ChatGPT in 2022, however, concern about malicious use of advanced ML has gone mainstream, to include bipartisan policymakers (King 2024), business leaders (Davis 2024), and the public (Orth and Bialik 2023). As these ML systems incorporate more modalities and are deployed more broadly, these concerns will likely remain significant for the foreseeable future.

These concerns do not exist in a vacuum—they motivate policy positions with significant implications. One example is whether to make models more or less accessible (Solaiman 2023). Some AI providers have cited malicious-use risks as reasons to opt for more controlled access (OpenAI 2023a). A second relevant debate centers on AI regulation. Policymakers and AI researchers often cite malicious-use risks as reasons why regulation of AI systems is necessary (Schumer 2023; Anderljung et al. 2023a; Bengio et al. 2023). A third debate is whether to halt development of more capable foundation models (Future of Life 2023). One justification for a ban is the lack of understanding of the malicious-use risks of today's systems (Aguirre 2023).

If concern about malicious use motivates important policy positions—including how to control AI models, how and how much regulation is warranted, and whether researchers should halt AI progress—then it is critical for researchers to rigorously unpack malicious-use risks. As Anderljung et al. (2023b) write, "good governance is in part an information problem." Specific research projects have been conducted in this vein. Researchers have tested whether language models can be used to generate propaganda articles (Kreps, McCain, and Brundage 2022; Goldstein et al. 2024), phishing emails (Hazell 2023), bioweapons (Soice et al. 2023; Mouton et al. 2024), and other types of hazardous content. Others have provided taxonomies and frameworks for categorizing risk (Bird, Ungless, and Kasirzadeh 2023; Blauth, Gstrein, and Zwitter 2022; Shevlane et al. 2023), harm (Hoffman and Frase 2023; Weidinger et al. 2022), and impact on society (Solaiman et al. 2023). We ask a related, but distinct question: How can researchers assess whether bad actors are likely to misuse a given AI system for particular malicious uses?

[*] Equal contribution. Authors listed in alphabetical order.

Our goal is to provide a simple, conceptual framework for assessing the likelihood of malicious use. We center this on what we call the "PPOu Framework." The PPOu Framework distinguishes three stages for assessing malicious use: Plausibility, Performance, and Observed use. The first, "Plausibility," examines whether a given system X can be used for a particular malicious use Y *at all*. Predeployment assessment at the plausibility stage often resembles red-teaming in cybersecurity, in that researchers adopt an adversarial mindset and try to use a system in a way unintended by the owner. The second stage, "Performance," asks *how well* a given system X can produce malicious use Y. Rather than assess whether a system can do something once (as in Stage 1), performance assessments help approximate how useful X is for a bad actor pursuing Y (and therefore how likely they are to use it). Relevant measurements include quality, robustness, cost-effectiveness, and improvement over other methods, including human baselines. The third and final stage, "Observed Use," seeks to determine whether and how system X is used for malicious use Y *in the real world*. There are selection effects when observing behavior that is designed to be secret, but observed use can shed light on malicious actors' intent, factor into deployment decisions, and update priors for what should be tested in the plausibility and performance stages.

Assessing the likelihood of malicious use is fundamentally an epistemic problem. Because models are trained, not hand designed, it is not always clear what capabilities a model has, or what level of effort or skill would be required to elicit the capabilities (Elhage et al. 2021). Similarly, the likelihood of malicious use is a function not only of the model, but also the threat actors' skills, resources, and intentions—which are difficult to ascertain. We describe strategies for reducing uncertainty for a predetermined set of Ys, but some level of uncertainty will remain.

The PPOu Framework serves at least four functions. First, it can help policymakers, academics, and researchers make sense of current strategies for assessing malicious-use risk. Policymakers who seek to develop regulations based on misuse should have a strong understanding of what types of data and benchmarks are feasible for these systems, and types of evidence that may be difficult to establish. Second, it can provide a resource for new AI labs and research groups for strategically testing new models prior to deployment (though in less detail than model cards [Mitchell et al. 2019]). Third, by centering methodologies, the framework provides entry points for a range of fields to contribute. Fourth, it provides a plug-and-play set of questions for empirically assessing newly identified or theorized risks.

In the remainder of this paper, we review existing strategies for assessing the likelihood of malicious use of advanced AI systems within the context of the PPOu Framework. In the following three sections, we describe how researchers assess malicious use at that stage, including key questions, methodologies, and strengths and limitations of those methods. We discuss the type of uncertainty addressed at each stage, and developments that would enable a more rigorous risk assessment than current approaches allow. The framework is intended to generalize to different malicious-use risks and to facilitate cross-domain learning in question formulation, study development, and policy discussion.

Of note, our focus is malicious use of advanced general-purpose systems—those with a potential for diverse economic and strategic applications—rather than AI systems that are designed for narrow harmful tasks. The risk of a foundation model producing hate speech falls into scope (Hartvigsen et al. 2022). An ML system designed specifically to identify faces of a certain type does not (e.g., Jain, Singh, and Vatsa 2018). We also focus on testing and evaluation of malicious use, rather than mitigations or prevention. Research into such efforts—such as building more robust guardrails for AI systems (see, e.g., Wallace et al. 2024) or ensuring the provenance of AI-generated content—is important, and should be informed by a nuanced consideration of misuse risks. Table 1 overviews the guiding question and methodologies for the three stages.

Stage	Guiding Question	Strategies to Reach an Answer
1. Plausibility	Is it plausible that system X could enable malicious use Y?	• Red-Teaming and Stress Testing • Probing by AI Models
2. Performance	How well can system X enable malicious use Y?	• Static Benchmarks • Lab, Survey, and Field Experiments • Modeling Marginal Utility to Bad Actors
3. Observed Use	Do people use system X for malicious use Y in the real world?	• Trust and Safety Monitoring • Investigations (OSINT and Journalism) • Analysis of Incident Databases

Table 1: Stages of the PPOu framework, with guiding questions and common methodologies.

Stage 0. Risk Identification

Before researchers can assess the risk of X for malicious use Y, they first need to identify X and Y. X will likely be fairly

obvious—a new system in training, or a product for potential customers. The set of Ys, however, is more difficult to ascertain.

In theory, there is an almost endless universe of possible malicious use applications of general-purpose AI systems. How can researchers identify which Ys they should apply the PPOu framework against? While a systematic review of threat ideation methodologies falls outside our scope (see Hicks et al. 2023, 26), a few ways to identify Ys include:

1. Asking experts in different domains.
2. Analyzing past incidents and misuses of related technologies (e.g., previous-generation systems).
3. Reviewing taxonomies of risks and drilling down into specific applications.
4. Considering coverage of policymaker concerns.
5. Using threat modeling methodologies to identify potential vulnerability points in high-risk systems.

In this paper, we describe applying the PPOu framework against a range of Ys that have been identified, recognizing that risk assessment is constrained by threat ideation.

Stage 1. Plausibility: Can X Be Used to Do Y?

Before AI labs open source models or release them to customers, there may be uncertainty about whether a given system (X) will likely be used for a certain malicious application (Y). Determining the plausibility that a system (X) can be used for some misuse case Y serves as the crucial first step in our framework. Can X do Y just once?

Conceptually, this involves either refining current understandings and predictions about the system, or formulating new hypotheses after hands-on interaction. For example, based on experience with previous AI systems, a researcher might speculate that X is capable of generating biological agents that require graduate-level knowledge. Historical capability forecasts can also inform hypotheses. In practice, this often means drawing on experts in the field of Y to apply X to that particular misuse case.

Consider whether a newly developed Large Language Model (LLM) can be used to engineer dangerous biological agents. Answering this question requires:

1. Understanding the (sub)skills required for Y, such as familiarity with dangerous biological agents.
2. Specifying what it would mean for X to "do" Y.
3. Attempting to use the LLM for this specific task.

Establishing plausibility does not require definitive proof that X can be used for Y. Rather, if testing demonstrates that X can accomplish tasks close to Y, this could be sufficient. For example, if an advanced AI system can generate novel chemical formulas that are not harmful, it might be plausible that it could also design harmful ones, even if it has not yet been applied to that task (e.g., Urbina et al. 2022).

Determining the threshold for what constitutes "plausible enough" is a judgment, as there is no standardized measurement. However, principles from safety-critical systems can offer guidance. To reduce the risk of underestimating capabilities of powerful AI systems, researchers could incorporate a safety margin when assessing capabilities (Eckert and Isaksson 2017). Post-training enhancements, like fine-tuning, can notably improve an AI system's capabilities (Davidson et al. 2023). The magnitude of these gains can inform a quantitative safety margin: if X does not seem capable of doing Y, but is close enough such that post-training enhancements could get it to do Y, then it may qualify as plausible. Phuong et al. discuss how this safety margin could be employed in practice (2024, section 6).

Common Stage 1 Methodologies

The fundamental difficulty with assessing plausibility is that whether a system can be used for a particular malicious use may be unknown ex ante. Reducing this uncertainty requires organizing experts to probe X for different malicious uses.

Red-Teaming and Stress Testing

The most common method for assessing plausibility of any given Y is red-teaming, or assigning individuals to think like an adversarial user to challenge one's systems or plans (Zenco 2015). In the context of cybersecurity, red-teamers are tasked with hacking into a system or network. In the context of AI systems, red-teamers "prob[e] AI systems and products for the identification of harmful capabilities, outputs, or infrastructure threats" (Frontier Model Forum 2023).[1] Red-teamers need not hack with code; they may "jailbreak" controls on models (Ji 2023) or test the use of models for dangerous applications. In other words, one of their goals is to uncover whether a system X can plausibly be used for Y. Experts can elicit specific behaviors by lightly optimizing the system to perform potentially malicious tasks, through improved prompting, tool-use enhancements, or scaffolding techniques.

The techniques employed by red-teamers are constrained by our scientific understanding of how ML produces highly capable systems. Because we do not yet have robust techniques for probing the internals of an AI system, red-teamers can only learn about the system's capabilities by modifying the input-output behavior (Casper et al. 2024). For example, a red-teamer with sufficient model access can triangulate

[1] A more precise definition is in the US Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. We acknowledge red-teaming is an imperfect term in the context of AI (Khlaaf 2023).

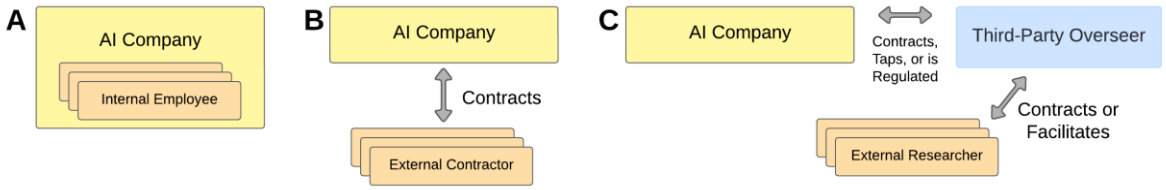

Figure 1: Three organizational structures for red-teaming.

dangerous capabilities of an AI system by fine-tuning it to do something dangerous (Phuong et al. 2024). This can increase confidence in assessments of the system's capabilities—if it cannot do Y even when trained to do so, then it is unlikely to be able to do Y by default. Similarly, a red-teamer might augment the system with biological design tools or custom software to help the system retry after failing on a task.[2] These techniques can help increase confidence in a plausibility assessment, even if it will still not be entirely conclusive.

Red teams vary in their organizational structure and their processes. Figure 1 illustrates three possible structures. In Model A, these processes could take the form of an AI company tasking employees with testing for malicious-use risk Y. Because the state of the art in these techniques is often contained within frontier AI companies, they have started employing specialized teams to explore these issues in-house (Anthropic 2024a). To the extent that red-teaming AMLSs requires expertise in those systems, in-house employees may be among the most effective plausibility testers. However, employees may lack expertise in certain threat vectors (OpenAI 2023b), be susceptible to groupthink (Janis 1997), or have bandwidth constraints. If employees believe their company seeks to ship products, they may be incentivized to underplay risks or not test for novel Ys.

Another organizational model is to hire individual external contractors (Model B). Outsourcing some of the red-teaming process has benefits (Anderljung et al. 2023b): external testers have a wider range of expertise, may be less influenced by organizational biases or company policies, and can serve as a public accountability mechanism by sounding the alarm if a technology is deployed prematurely or without adequate oversight. Companies could innovate by running bug bounty programs similar to those in cyber, rewarding contractors who discover the most pernicious risks (Kuehn and Mueller 2014.).

A salient concern for Model B is the possibility of industry capture. To protect the organization's IP, external red-teamers will often need to sign a nondisclosure agreement. Researchers should explore whether NDAs are a deterrent for speaking publicly about risks of X for Ys if dangerous capabilities exceed a threshold or labs mischaracterize red-teaming results. Longpre et al. (2024) argue that researchers would benefit from legal and technical safe harbor (e.g., protections for good faith research and from account removal). Another concern is that red-teaming could be designed to assuage regulators (Feffer, Sinha, and Lipton 2024), as self-regulation can undermine public support for more stringent public regulation (Malhotra, Monin, and Tomz 2019).

Another approach would be for an external third party to oversee red-teaming. An AI company could contract or tap a third-party overseer (self-regulation). However, legislation could also mandate that a third-party overseer (e.g., an AI Safety Institute) oversee the process and contract or facilitate access to external researchers (Model C).[3] A single actor need not be responsible for all types of testing: a small number of tests could be mandated and carried out by government (e.g., under Model C) while a broader number of lower-risk tests could be overseen by companies (e.g., under Model A or B) (Anthropic 2024b).

An independent body overseeing AI red-teaming not only would increase perceived objectivity of the process, but also could capitalize on synergies. For example, if companies run independent processes, researchers who red-team for three companies get vetted thrice. A third-party body—whether the industry-led Frontier Model Forum or a government-led AI Safety Institute—could provide one check. Such a body could develop best practices for how to share models, under what conditions NDAs are required, and how to deal with information hazards. They could even run testing themselves, akin to the independent Insurance Institute for Highway Safety's crash testing of cars.[4] These efforts could increase assurance that red teams maximize their likelihood of X producing Y, and that the plausibility finding is made public as necessary.

Machine Red-Teaming

While the pool of red-teamers is increasing, the labor market for human red-teaming is necessarily limited in size. In parallel, AI systems are improving at a rapid rate. Applying ma-

[2] See Zhong, Wang, and Shang (2024) in the context of engineering.

[3] If testing is mandated, it will be critical that the agency responsible for oversight has the resources to do so, to avoid an orange book situation. See Department of Defense (1985). On regulatory markets for AI safety, see Clark and Hadfield (2019).

[4] See https://www.iihs.org/.

chines to red-team machines is a technique that both naturally benefits from AI improvements and can, in principle, flexibly expand the red-teaming labor force.

There are several reasons that machine red-teaming is technically promising. First, it is often easier to evaluate an output than it is to generate it. Machine learning researchers have appealed to this regularity to design schemes to supervise models that exceed human performance. Similarly, this regularity can potentially allow for training machine red-teamers that evaluate outputs of a target model. Second, machine red-teaming can complement humans by taking advantage of machine advantages, such as cheaply scaling content generation. For example, consider red-teaming for cyber-vulnerabilities, which can require systematic and tedious generation of test cases. Systems trained to write code may be able to more quickly and flexibly generate examples than human coders.

Automated red-teaming is a relatively nascent field, but there are early signs that even today's AI systems can be leveraged for red-teaming. For example, one paper found that current AI systems can be used to identify social groups that a target AI system discusses in offensive ways (Perez et al. 2022). Another paper found that models can produce critiques that help humans identify flaws in model outputs (Saunders et al. 2022). As we discuss below, this methodology offers a promising avenue for investment.

Discussion of Stage 1. Plausibility

Assessing malicious use at the plausibility stage has some clear advantages. Most notably, plausibility testing can be conducted prior to an AI system's release. This allows developers to begin assessing risks before real people are subjected. Red-teaming exercises can also be conducted more quickly than research at the performance stage and involve a wide range of testers.

However, there are also a slew of limitations (e.g., Walsh 2024). A particularly pernicious problem is what we call the *capability elicitation problem*. AI systems often develop unexpected capabilities, because they are general purpose and the science of capability prediction is not that advanced yet. These capabilities are unearthed through much trial and error. But if the judgment is that X cannot do Y (or anything close to Y), how do we know that's actually the case? Perhaps an improved prompt (a recent paper found that prepending "Take a deep breath" before the requested task improved performance [Chengrun et al. 2023]) could cause the system to generate credible ideas for novel and harmful biological agents. Equipping the model with tools (like biological design tools) or giving it the ability to retry a task could also elicit unforeseen capabilities.

If the dominant factor in plausibility testing is the sheer scale of labor involved, then it may be difficult to determine when *enough* red-teaming has taken place such that public experimentation will be unlikely to unveil novel dangerous capabilities for the Ys considered. If companies red-team at small scale, they may have false assurances that systems cannot be misused in ways that are later discovered by the public. Improved techniques for capability elicitation could have a large impact (Greenblatt et al. 2024), and researchers are working on methods like automated techniques to find the best prompt for a given task (Deng et al. 2022). Today, however, state-of-the-art capability elicitation requires extensive experience interacting with the AI system. Furthermore, even if machine-augmented red-teaming becomes the norm, the capability elicitation problem will still apply.

Another set of limitations stems from the red-teaming process. Red-teaming may not generalize to other models because of differences in training or guardrails. Results from a red-team exercise today may also not translate to future models that are more capable; frequent testing may become costly. (For a costing data point, see Ganguli et al. 2022.) Processes also face a trade-off in granting greater access: bringing in more researchers to the red-teaming process could better uncover malicious-use risks, but may also carry greater risks if approved researchers inappropriately share access with others (akin to the Cambridge Analytica case) (Rehman 2019). In niche domains, subject matter experts can be difficult to find and evaluate. Finally, there may be inherent risks of red-teaming itself, analogous to "gain of function" concerns in biotechnology. If a red-teamer is attempting to get a model to do something dangerous by fine-tuning it to be dangerous, then that itself could be risky if appropriate safeguards (e.g., preventing wider distribution of the fine-tuned model) are not in place.

To scale up red-teaming processes to match potential growth in new models, we need better institutional infrastructure. While AI firms currently use small teams of external red-teamers, we can imagine an industry of "AI auditors" that provide red-teaming as a service. Competition among competent red-teamers might increase the quality of plausibility testing services. Finally, the science of evaluating these systems will need to improve. Better methods for gaining assurance that the model does not contain the target behavior will help reduce the capability elicitation problem.

Stage 2. Performance: How Well Can X Do Y?

Stage 1 aims to reduce uncertainty about misuse risks by testing whether a system X can produce Y *at all*. If an AI lab is concerned about whether X could be used to persuade people of a viewpoint, the first step would be to test whether the system can even express that viewpoint. An affirmative answer only partially reduces uncertainty about the threat vector. A high degree of uncertainty would likely still remain about whether a threat actor could benefit from using

X for Y (and thus might be likely to do so). This calculation centers on factors such as how well, reliably, cost-effectively, and easily X can do Y. Methodologies at Stage 2 address these follow-on questions. Just as social scientists test hypotheses about human or animal behavior in controlled settings, researchers at the performance stage sample behaviors from X and evaluate them (via humans or models), ideally in a systematic and reproducible way.

Common Stage 2 Methodologies

State 2 methodologies include testing AI systems against static benchmarks, running experiments (lab, survey, or field), and conducting marginal utility analysis (such as cost modeling). We unpack each of these methods.

Static Benchmarks

To gain evidence about these different factors of performance, researchers employ a portfolio of methods. First, there are the equivalent of written tests. These benchmarks test the ML system against a known standard or set of questions. A researcher might reason that for an ML system to be useful in producing biological agents, it will need some knowledge and subskills from biology. They can then test the system on a benchmark such as a graduate biology exam.

Static benchmarks are common in machine learning. For example, benchmarks have been constructed to test language and image understanding, graduate-level problem-solving, and legal reasoning (see, e.g., Paper with Code's "Browse State-of-the-Art"). Static benchmarks can provide a standardized and relatively low-cost method to assess different AI systems in a reproducible way. As many static benchmarks have been created, researchers create indices that weight individual benchmarks differently to get a more holistic sense of an AI system's capabilities in a domain (see, e.g., the HELM leaderboard). Their low cost can help static benchmarks serve as cheap proxies for more expensive and higher-fidelity methods of evaluation, like data from observed use.

However, static benchmarks face limitations. Practically, compelling static benchmarks are notoriously difficult to create. Without careful construction, the benchmarks are often polluted—the ML system has already seen some of the answers during training. ML systems may take advantage of subtle biases in the benchmark questions (e.g., if the correct answer on a multiple choice exam is the longest one). Researchers may also optimize a model for high performance on benchmarks and skew an assessment of true capabilities (over-fitting). For certain classes of malicious use (e.g., highly specialized and controlled domain knowledge, like nuclear science), it may be difficult to create static benchmarks that a wide range of actors can test against. Finally, results on static benchmarks have limited external validity, since real-world performance is often context dependent and requires adapting to changing conditions.

Lab, Survey, and Field Experiments

Researchers can run experiments in which they deliberately introduce an AI system or AI-generated content and determine the effect on outcomes of interest—from the utility of AI in coding (Poldrack, Lu, and Beguš 2023) to the ability of LLMs to persuade people not to believe conspiracies (Costello, Pennycook, and Rand 2024). There are a variety of different settings—including lab, survey, and field—where researchers could assess not only positive effects of AI, but malicious-use risks as well.

In a lab or survey experiment, researchers can assess the causal effect of a treatment in a tightly controlled setting (Falk and Heckman 2009). For example, Spitale et al. (2023) showed participants GPT-3-generated fake tweets and disinformation tweets written by Twitter users and found that humans were worse at identifying AI-generated disinformation compared to false human content. In a laboratory environment, researchers can assess factors like ease of use. For example, researchers can equip an undergraduate student with an LLM and access to a biolab, and see if the student can then manufacture a biological agent.

Lab and survey experiments have advantages: researchers can run different treatments, recruit respondents using existing pools (e.g., through providers such as Lucid), and ensure informed consent. However, for assessing malicious-use risks, researchers will still face limitations because of the duty to minimize harm to respondents (and meet the criteria of their IRB) and the difficulty in generalizing from the lab to the real world.

In a field experiment, researchers assess the efficacy of an intervention in the wild (Baldassarri and Abascal 2017). Governments used large-scale field experiments in the late twentieth century to test the efficacy of government programs (Levitt and List 2009), and researchers in the social sciences have used the method to test interpersonal and political behaviors from racial discrimination in hiring (Quillian and Midtbøen 2021) to the effect of campaign behavior on voting (Kalla and Broockman 2018).

Researchers can sometimes devise ways to test how well X can do Y in a natural environment. For example, Kreps and Kriner (2023) examined the risk of a bad actor using a language model to falsify constituent sentiment. They sent letters written by humans as well as AI-generated letters to 7,132 US state legislators and measured the rate and length of response from the offices (Kreps and Kriner 2023). The study found that legislators responded to AI-written constituent letters only a few percent less frequently than human-written ones—suggesting a malicious-use risk if bad actors pursued a similar strategy.

In addition to the limitations of field experiments in general (Baldassarri and Abascal 2017), field experiments around malicious use may be heavily constrained by practical and ethical considerations. To be maximally informative, experiments would require bad actor collaboration—

researchers can not directly study how the Russia-backed Internet Research Agency (IRA) would use a new AI system for propaganda without equipping the IRA with that AI system. This, much like deploying an AI-designed biological agent to evaluate if it works, would clearly be unethical. Researchers could design prosocial field experiments (e.g., working with companies to integrate an experiment about AI-generated phishing into pre-planned phishing email training), but some Ys are simply not ethical to test on the public. Experiments involving human subjects are often more expensive to execute than static benchmarks.

While experiments represent a promising direction, on some tasks, human baselines may be insufficient and/or alternative machine baselines must be considered. Studies have found that GPT-4 can persuade people in debate as well as humans can (Salvi 2024), and persuasive capabilities increase with model size (Durmus et al. 2024). If language models become as persuasive as the best human interlocutors, it may be necessary to compare results of persuasion experiments across the results of other models (rather than humans). Likewise, even if a system demonstrates effectiveness at a set of reproducible benchmarks, that does not imply that bad actors will judge that it is worth using. If a system can reliably produce instructions for designing a bioweapon, but so can a simple Google search, then the marginal utility to bad actors is limited (Scott et al. 2024).

Modeling Marginal Utility to Bad Actors

Instead, we might assess if X provides an uplift over other alternative tools for the same task. If it doesn't, then bad actors are unlikely to choose X for their malicious use (Kapoor et al. 2024). Again, the debate about the biosecurity risks of ML systems is relevant here. In June 2023, a preprint paper detailed an exercise in which MIT students were given access to a chatbot to test whether they could use it to suggest new pandemic-class pathogens. The students were able to get the ML system to suggest four potential pandemic pathogens and to give instructions on how to build these (Soice et al 2023). The paper gained traction among policymakers and media.

But the study did not assess whether chatbots outperformed other tools, like Google. Drawing any conclusions about the biosecurity risk from ML systems would be premature. Follow-up studies that assessed uplift led to a messier picture of the threat (Mouton, Lucas, and Guest 2024).

These uplift tests require understanding the state-of-the-art in current tools and how they are used. This often requires human experts in risk domains of interest. For example, to assess malicious uses in cybersecurity, we can compare the uplift from X with:

1. Specialized software tools, e.g., fuzzers that attempt to automatically generate inputs that trigger security vulnerabilities.
2. Single humans with these specialized tools.
3. Large human-machine teams with specialized tools.

Depending on the X and Y, some comparisons may be more relevant than others. If the concern is an ML system *autonomously* carrying out cyberattacks, then it makes more sense to compare it to human experts. If the concern is whether X is better at a specific capability like fuzzing, then it makes more sense to compare X to existing software tools like AFL++.

Other factors, like the system's reliability, cost-effectiveness, and ease of use at the task are also a key part of the calculus for bad actors. For example, if X is extremely expensive and finicky, only to produce malware that is no better than the current non-X state-of-the-art, then the likelihood of the threat is likely small, even if malicious use is technically feasible. One study conducted a cost modeling analysis of using an LLM for an influence operation and assessed how the percentage of usable outputs and the likelihood of discovery impact the cost savings for propagandists (Musser 2023). Studies could model the relative utility of other malicious uses under different conditions.

Discussion of Stage 2. Performance

The most salient benefit of performance testing mirrors the benefits of good science, by systematically and skeptically testing hypotheses that emerged from the plausibility stage. Showing that you can do something once (plausibility) does not mean it is a useful tool for doing it. More subtly, performance testing can help refine earlier hypotheses about malicious use by reshaping the Ys that are worth prioritizing. Suppose that red-teamers find it plausible that an AI system could aid a human in exploiting software vulnerabilities. The performance stage may find that the system is actually best at exploiting a particular class of web application vulnerabilities, leading to further plausibility tests.

The other key benefit of the performance stage is its role in assessing marginal risk introduced if bad actors have access to the ML system in the real world. This assessment hinges on factors like ease of use, quality, reliability, cost-effectiveness, and uplift over existing methods. Evaluating these aspects helps stakeholders make informed decisions by balancing the system's benefits against its risks. For example, a system's ease of use and reliability can significantly influence its practical impact and the likelihood of adoption. Similarly, its cost-effectiveness affects accessibility and potential scale of use.

Performance testing is limited in how well its results can generalize to the real world. At best, static benchmarks and production experiments are proxies for real-world performance. This limits how much researchers can learn about malicious-use risks from the performance stage. Bad actors may also have access to different variants of the ML system,

such as a version without guardrails. Thus, it is likely important to assess the performance of different variants of X.

How could performance testing evolve over time? Continued technical progress could involve searching for scaling analysis for malicious use. These are simple mathematical relationships that allow a researcher to quantitatively forecast the behavior of more powerful ML systems.[5] While it may not be possible to find literal laws for malicious use, investigating the scaling properties of malicious use could reduce uncertainty about the riskiness of future systems. Other technical progress could involve developing better techniques to help generalize results from the lab to the field. Perhaps researchers could use metrics from the observed use stage as a proxy for actual outcomes, which could be leveraged in the performance stage. From a research design standpoint, estimating likelihood of use is not only contingent on performance of the model, but also on defenses or mitigations that could function to neutralize or deter malicious use. Lab or field experiments that compare performance of AI models to relevant benchmarks may not necessarily take mitigations into place, but for some abuse types this may be feasible.

Institutional progress could involve developing arrangements that minimize the incentive issues at this stage. Independent third-party performance testing can help impartial observers evaluate different Xs. Online leaderboards do this for general AI capabilities,[6] and a number of government-sponsored and independent organizations are cropping up to do this for safety and misuse. But there are still gaps. For one, as Xs get more powerful, these organizations will have to be held to stringent information security standards—a cost that could be more bearable by better-resourced actors, like frontier AI labs.

Another incentive issue is more subtle. AI companies might have the best capability elicitation and testing techniques because of their accumulated experience and ML expertise (see, e.g., Day and Montgomery 1983). If some Y is only reliable after fine-tuning the model, then that fine-tuning is necessary to evaluate whether X can do Y well. But AI companies might face incentives (e.g., pressure to go to market, resource allocation trade-offs) to sandbag or otherwise not thoroughly test their own systems. In the future, alleviating this issue might require on-site third-party auditors or sharing proprietary tools with third-party assessors.

Stage 3. Observed Use: Do People Use X To Do Y?

The two stages described above—plausibility and performance—primarily focus on machine capability: whether, and how well, a given system can perform a function of interest. Research prior to deployment can estimate adversarial intentions, but researcher expectations and actual misuse will likely differ.

Misestimation could take place for reasons from cognitive biases such as mirroring (assuming adversaries will misuse systems in ways analysts themselves would) to unknown capabilities (emergent abilities that were not known at release date are then used for malicious purposes).[7] Misestimations in security domains are routine. For example, in the 1980s, US government analysts were concerned that if source code for encryption was public, the Soviets would break into US networks. That was not the case; rather, the threat that emerged was criminals using encryption to evade detection of bad behavior (like child sexual abuse material). The third step of the framework moves from capabilities of the system to whether and how people use system X for malicious use Y in the real world—its observed use.

Three factors can guide analysis at this stage: *application*, *actor*, and *frequency*. To illustrate the analytical usefulness of these factors, we'll consider a scenario where researchers conduct plausibility and performance testing to evaluate the risk of rogue terrorist groups using LLMs to devise dangerous chemical mixtures.

The first factor is *application*: What is the specific embodiment of misuse? At the plausibility stage, researchers may decide that a risk seems plausible: rogue terrorists could use an AI system to help create novel chemical compounds to build an extremely destructive weapon. But imagine that, in practice, rogue terrorists use an LLM not for instructions for creating a new chemical but rather for the logistical process of successfully carrying out a terrorist attack (e.g., determining timing, location, best practices for communication among team members, etc.). Knowing the nature of the specific use can help update threat models and surface needed mitigations.

The second factor is *actor*: Who is using system X for malicious use Y? Researchers may expect that rogue terrorist groups might use a language model for ideas to create chemical weapons; in practice, they may find that a different actor type, like a state with low knowledge or resources in chemistry, actually does so. Understanding which actors are misusing in practice is useful for updating the threat model and for mitigation development (in this case, a mitigation

[5] These "scaling laws" generally apply to fundamental measures of performance like the predictive loss for generative models (Kaplan et al. 2020). However, early work characterizes scaling laws for more specific behaviors (e.g., Ghorbani et al. 2021 for machine translation).

[6] See, e.g., https://chat.lmsys.org/.

[7] Progress on the capability elicitation question would help reduce the risk of misestimation.

could be nation-state pressure). Of course, it may not be possible to determine the identity of bad actors; research describes challenges with making attribution and shows that attribution can sometimes be made to certain levels but not others (e.g., identifying an IP address that is the source of an attack, but not the human behind the computer) (Lin 2016).

Finally, the third factor is *frequency*: How often is system X used for malicious behavior Y? If a single terrorist is caught asking a model how to create a chemical weapon one time, the malicious-use risk is very different than if dozens of groups routinely ask the model this question or follow-up questions that suggest subsequent offline action.

Importantly, observing malicious use can also inform questions about the *severity of harm*: how grave is the observed malicious use? Assessments of harm shape important decisions, including how much user friction to prevent malicious use Y is warranted and how to weigh malicious use Y against other risks. This information can be useful to identify potential malicious-use applications (which Ys?) that are worth testing in future risk assessments (see, Stage 0).

Common Stage 3 Methodologies

Answering these questions requires investigating malicious use in the real world. A variety of different methodologies could surface this behavior, including monitoring from trust and safety teams at AI labs, open-source intelligence (OSINT) methods and investigative journalism, and trend analysis of incident databases (Goldstein and Lohn 2023).

Trust and Safety Monitoring

If a given advanced AI system is controlled by a private actor, that opens up a direct route for discovery: ongoing monitoring. Companies that make systems available to developers could monitor API calls; those that create usable interfaces for users could analyze prompts and responses. Monitoring regimes need not treat all users equally: for example, there could be differential monitoring based on application type, user type (e.g., verified users versus suspicious users), or even time period (e.g., around an election). Monitoring by AI providers may pose safety-privacy trade-offs analogous to monitoring by social media platforms: while greater oversight can surface bad behavior, it may also infringe on user privacy and user expectations. Privacy-preserving practices in monitoring are important to alleviate this trade-off.

Investigations: OSINT and Journalism

However, many advanced AI systems could be distributed openly (e.g., with open, downloadable weights). In this case, investigators and researchers can use OSINT methods (Pastor-Galindo et al. 2020) to surface cases of malicious uses. OSINT can be surprisingly effective: It has been used to spot criminal intentions or hold perpetrators to account, track cyberattacks (Anderson and Guarnieri 2017), document war crimes (Bellingcat Investigative Team 2022), and more.

Detecting malicious use of AMLSs could occur without directly observing the behavior or content produced. Indications that bad actors are using AMLSs in their workflows may be useful. Journalists can interview actors to ask them about whether and how they use AMLSs, monitor discussions on criminal forums, leaks of chat logs or purchasing data from known bad actors to AI providers, and more (Lin et al. 2024). Journalists are often leaders in exposing operational behavior and harm from malicious actors. For example, Clarissa Ward at CNN unveiled behaviors of Ghanaian disinformation-for-hire employees—e.g., they worked from phones, rather than laptops, shared a rent-free apartment, etc.—that would be difficult to ascertain from technical investigations alone (Ward et al. 2020).

Analysis of Incident Databases

Rather than study single cases of malicious use, a path to better understand the abuse may be to develop databases for comparative and aggregate analysis. Industries as diverse as healthcare and aviation have incident reporting databases, which often capture three types of incidents: adverse events, "no harm events," and "near misses" (Wald and Shojania 2001). Incident reporting typically focuses on accidents (or near accidents), which is distinct from the question of this paper of intentional misuse. Nevertheless, the incident reporting framework may be useful for providing lessons for recording observed use. For example, voluntary versus required disclosure may produce different rates of inclusion; dedicated funding to employees who aggregate content could ensure such a database is as exhaustive as possible given public knowledge; and uniform classification across inputs may better foster analysis (Dixon and Frase 2024). So long as incident reporting is voluntary, it should be incentivized by allowing for anonymous reporting to a trusted third-party body (Avin et al. 2021).

Discussion of Stage 3. Observed Use

If the goal is to understand the likelihood of malicious use of AMLSs, observed use offers direct evidence. Unlike the plausibility and performance stages, observing malicious use does not require making assumptions about what certain actors or actor types might do. It also can provide leads for additional research, where similar actors or actor types face similar incentives or constraints. Findings from observed use can also inform the identification of misuses for plausibility and building of benchmarks for performance.

However, we expect at least three limitations. The first is data *completeness* and *representativeness*. Similar to the challenge of creating reliable datasets about cyberattacks, many malicious uses of advanced ML systems are likely to go undetected, making it difficult to assess harm—even at the observed use stage. Furthermore, if the undetected cases are systematically different from the detected ones, then the resulting analysis will be biased (from a selection effect).

Researchers' assessments based on cybersecurity incident databases suffer from systematic bias, because the universe of possible cases is not observed (Baram, Vićić, and Gartzke 2023). As Baram et al. point out, the ostensible solution of simply collecting more data may actually attenuate existing biases if nonpublic cases differ from public ones.

The second limitation is *ethical*: If the populations that may be harmed by models are not included in decision making about system release, then AI providers may not be respecting their autonomy. Who should make release decisions and the principles that should guide those processes are not normatively settled.

Third, and finally, is *projection*: Observed use of system X today may not generalize to other systems or to the future because of changes in intentions or capabilities. Potential threat actors could start adopting tools even if they did not use earlier generations, and newer Xs could be useful for Y even if previous Xs were not.

Bolstering the usefulness of the Stage 3 (Observed use) requires progress on a few fronts. The first is capacity to monitor. While AI providers employ malicious-use classifiers and have teams dedicated to monitoring and enforcing these policies, there is not yet third-party verification that AI companies will actually catch particular misuses. Similarly, research into privacy-preserving practices for monitoring could alleviate trade-offs between privacy and observing misuse. Technical work on strengthening malicious-use classifiers could also reduce administrative costs and make scaled monitoring more feasible.

More reliable observational data on the distribution of outputs could further help reduce uncertainty at this stage. Consider influence operations that are conducted on an AI API. It is often hard to determine whether any single output is malicious without knowing how it is distributed to and affecting real people.

One method to address the *projection* issue—whether and when future Xs can be used for Y—is judgment-based forecasting. This could take the form of large-N surveys of AI and ML researchers to better estimate progress (Zhang et al. 2022; Grace et al. 2018), or the Delphi method, in which a group of experts are asked their opinions of an issue, updating their independent views based on others' strength of reasoning (Barrett and Heale 2020). Research on forecasting shows that forecaster calibration and training can improve results (Zellner et al. 2021).

Judgment-based forecasting faces its challenges, too. The track record on AI questions is mixed: some forecasts by experts of progress, such as on the MATH benchmark, were significantly behind actual progress by 2023 (Steinhardt 2022). A review of 400 retrospective, long-term forecasts from 16 papers found most papers had a success rate of less than 50% on forecasts made (Mullins 2018). Even if best practices are followed, a high degree of uncertainty about future malicious-use capabilities may remain.

Conclusion

As AI systems become more advanced and accessible, investigating misuse becomes increasingly important. This paper provides a framework for how researchers assess the likelihood of malicious use through three sequential stages: (1) Plausibility (can X do Y at all?) (2) Performance (how well can X do Y?) and (3) Observed use (do people use X to do Y in practice?) These three stages also attempt to reduce different types of uncertainty. Stage 1 reduces uncertainty about feasibility, but leaves gaps about the utility of a system compared to alternatives. Stage 2 reduces uncertainty further by testing the marginal uplift, cost-effectiveness, reliability, and robustness of systems. However, this still leaves uncertainty about interactions in the real world, such as which actors will attempt to misuse AI systems, and how their attempts will interact with defensive measures. Assembling real-world use cases, Stage 3, attempts to fill this hole. Thus, policymakers should not focus exclusively on any one stage as some uncertainty will always remain.

Our framework faces limitations about breadth, linearity, and uncertainty. First, the PPOu framework is a tool for exploring whether or not—and how—system X is likely to result in malicious use Y, where Y is *some particular malicious use*. If no one asks "Can an AI system with voice synthesis capabilities and a radio-access agent intrude into FAA radio channels and cause catastrophic accidents?", then researchers may not attempt to red-team for relevant capabilities nor conduct performance experiments to test whether pilots would be fooled. The PPOu framework can only reduce uncertainty for the Ys it is applied against. Similarly, we mainly discuss human uses of X, rather than automated uses of X (as in the case of agentic systems).

Second, our framework is more linear than the iterative process for assessing malicious use in the real world. Because performance testing is time-consuming, such efforts may occur alongside observed use investigations. Likewise, observed misuse could lead to new rounds of red-teaming that were not conducted prior to model release, as red-teamers had not previously identified the Y.

Finally, some degree of uncertainty about the likelihood of using X for Y will always remain. The capability elicitation problem highlights that it may not be possible to know whether X can produce Y with complete certainty. Some forms of production experiments would not be ethical to run, so uncertainty about how a system would perform at Y may have to remain. Observed use can help update prior knowledge, but covert activity will still have falsifiability challenges. The goal of this framework—and the broader field of risk assessments on which we build—is not to reduce uncertainty *altogether*, but to do so *to the extent possible* (for Ys that have been conceived of).

Acknowledgments

For feedback on an earlier draft of this paper, we thank Lama Ahmad, Markus Anderljung, John Bansemer, Rosie Campbell, Derek Chong, Jessica Ji, Igor Mikolic-Torreira, Andrew Reddie, Abhiram Reddy, Chris Rohlf, Colin Shea-Blymyer, Weiyan Shi, Toby Shevlane, Thomas Woodside, and participants at the NLP SoDaS Conference 2023. For editing support, we thank Eden Beck. All errors remain our own.

References

Aguirre, A. 2023. Close the Gates to an Inhuman Future: How and Why We Should Choose to Not Develop Superhuman General-Purpose Artificial Intelligence. arXiv preprint. arXiv:2311.09452 [cs.CY]. Ithaca, NY: Cornell University Library. doi.org/10.48550/arXiv.2311.09452.

Anderljung, M.; Barnhart, J.; Korinek, A.; Leung, J.; O'Keefe, C.; Whittlestone, J.; Avin, S.; Brundage, M.; Bullock, J.; Cass-Beggs, D.; et al. 2023a. Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv preprint. arXiv:2307.03718v4 [cs.CY]. Ithaca, NY: Cornell University Library.

Anderljung, M.; Smith, E. T.; O'Brien, J.; Soder, L.; Bucknall, B.; Bluemke, E.; Schuett, J.; Trager, R.; Strahm, L.; and Chowdhury, R. 2023b. Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework. arXiv preprint. arXiv:2311.14711 [cs.CY]. Ithaca, NY: Cornell University Library.

Anderson, C. and Guarnieri, C. 2017. Bahamut, Pursuing a Cyber Espionage Actor in the Middle East. Amsterdam, Netherlands: Bellingcat. https://www.bellingcat.com/news/mena/2017/06/12/bahamut-pursuing-cyber-espionage-actor-middle-east/. Accessed: 2024-04-29.

Anthropic. 2024a. Job Posting: Research Scientist, Frontier Red Team (Cyber). https://jobs.lever.co/Anthropic/8f565d59-8831-443a-b72a-cb9ef8ae06b2. Accessed: 2024-04-29.

Anthropic. 2024b. Third-Party Testing as a Key Ingredient of AI Policy. https://www.anthropic.com/news/third-party-testing. Accessed: 2024-04-29.

Avin, S.; Belfield, H.; Brundage, M.; Krueger, G.; Wang, J.; Weller, A.; Anderljung, M.; Krawczuk, I.; Krueger, D.; Lebensold, J.; Maharaj, T.; and Zilberman, N. 2021. Filling Gaps in Trustworthy Development of AI. *Science* 374(6573): 1327–1329. doi.org/10.1126/science.abi7176.

Baldassarri, D., and Abascal, M. 2017. Field Experiments across the Social Sciences. *Annual Review of Sociology* 43: 41–73. doi.org/10.1146/annurev-soc-073014-112445.

Baram, G.; Vićić, J.; and Gartzke, E. 2023. What Lurks Beneath the Tip of the Iceberg? Exploring the "Missingness" Problem in Cyber Events Data. Paper presented at the American Political Science Association Annual Convention. Los Angeles, August 31–September 3.

Barrett, D., and Heale, R. 2020. What Are Delphi Studies? *Evidence Based Nursing* 23(3): 68–69. doi.org/10.1136/ebnurs-2020-103303.

Bellingcat Investigative Team. 2024. Civilian Harm in Ukraine. https://ukraine.bellingcat.com/. Accessed: 2024-04-29.

Bellingcat Investigative Team. 2022. Tracking the Faceless Killers Who Mutilated and Executed a Ukrainian POW. Bellingcat. https://www.bellingcat.com/news/2022/08/05/tracking-the-faceless-killers-who-mutilated-and-executed-a-ukrainian-pow/. Accessed: 2024-5-13.

Bender, E. M.; Gebru T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery. doi.org/10.1145/3442188.3445922.

Bengio, Y.; Hinton, G.; Yao, A.; Song, D.; Abbeel, P.; Harari, Y. N.; Zhang, Y.-Q.; Xue, L.; Shalev-Shwartz, S.; Hadfield, H.; Clune, J.; Maharaj, T.; Hutter, F.; Baydin, A. G.; McIlraith, S.; Gao, Q.; Acharya, A.; Krueger, D.; Dragan, A.; Torr, P.; Russell, S.; Kahnemanm, D.; Brauner, J.; and Mindermann, S. 2023. Managing AI Risks in an Era of Rapid Progress. arXiv preprint. arXiv:2310.17688 [cs.CY]. Ithaca, NY: Cornell University Library.

Bird, C.; Ungless, E.; and Kasirzadeh, A. 2023. Typology of Risks of Generative Text-To-Image Models. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. New York: Association for Computing Machinery. doi.org/10.1145/3600211.3604722.

Blauth, T.F.; Gstrein, O.J.; and Zwitter, A., 2022. Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI. In *IEEE Access* 10: 77110–77122. doi.or/10.1109/ACCESS.2022.3191790.

Bommasani, R.; Kapoor, S.; Klyman, K.; Longpre, S.; Ramaswami, A.; Zhang, D.; Schaake, M.; Ho, D. E.; Narayanan, A.; and Liang, P. 2023. *Considerations for Governing Open Foundation Models*. Foundation Model Issue Brief Series. Stanford, CA: Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf.

Brundage, M.; Avin, S.; Clark, J.; Toner H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, S.; Zeitzoff, T.; Filar, B.; et al. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv preprint. arXiv:1802.07228 [cs.AI]. Ithaca, NY: Cornell University Library.

Casper, S.; Ezell, C.; Siegmann, C.; Kolt, N.; Curtis, T. L.; Bucknall, B.; Haupt, A.; Wei, K.; Scheurer, J.; Hobbhahn, M.; Le Sharkey, L.; Krishna, S.; Von Hagen, M.; Alberti, A.; Chan, A.; Sun, Q.; Gerovitch, M.; Bau, D.; Tegmark, M.; Krueger, D.; and Hadfield-Menell, D. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. arXiv preprint. arXiv:2401.14446v1 [cs.CY]. Ithaca, NY: Cornell University Library.

Chengrun, Y.; Wang, X.; Lu, Y.; Liu, H.; Le, Q.V.; Zhou, D.; and Chen, X. 2023. Large Language Models as Optimizers. arXiv preprint. arXiv:2309.03409 [cs.LG]. Ithaca, NY: Cornell University Library.

Clark, J., and Hadfield, G. K. 2019. Regulatory Markets for AI Safety. arXiv preprint. arXiv:2001.00078v1 [cs.CY]. Ithaca, NY: Cornell University Library.

Costello, T. H.; Pennycook, G.; and Rand, D. 2024. Durably Reducing Conspiracy Beliefs through Dialogues with AI. PsyArxiv preprint. doi.org/10.31234/osf.io/xcwdn.

Davidson, T.; Denain, J-S.; Villalobos, P.; and Bash, G. 2023. AI Capabilities Can Be Significantly Improved without Expensive Retraining. arXiv preprint. arXiv:2312.07413v1 [cs.AI]. Ithaca, NY: Cornell University Library.

Davis, J. 2024. What Business Leaders Really Think about Generative AI. INSEAD. https://knowledge.insead.edu/leadership-organisations/what-business-leaders-really-think-about-generative-ai#:~:text=Concerns%20about%20outsmarting%20humans%20. Accessed: 2024-5-2.

Day, G. S., and Montgomery, D. B. 1983. Diagnosing the Experience Curve. *Journal of Marketing* 47(2): 44–58. doi.org/10.2307/1251492.

Deng, M.; Wang, J.; Hsieh, C.-P.; Wang, Y.; Guo, H.; Shu, T.; Song, M.; Xing, E. P.; and Hu, Z. 2022. RLPROMPT: Optimizing Discrete Text Prompts with Reinforcement Learning. arXiv preprint. arXiv:2205.12548v3 [cs.CL]. Ithaca, NY: Cornell University Library.

Department of Defense. 1985. Trusted Computer System Evaluation Criteria (DoD 5200.28-STD). https://csrc.nist.gov/files/pubs/conference/1998/10/08/proceedings-of-the-21st-nissc-1998/final/docs/early-cs-papers/dod85.pdf.

Dixon, R. B. L., and Frase, H. 2024. *An Argument for Hybrid AI Incident Reporting*. Washington, DC: Center for Security and Emerging Technology. doi.org/10.51593/20230046.

Durmus, E.; Lovitt, L.; Tamkin, A.; Ritchie, S.; Clark, J.; and Ganguli, D. 2024. Measuring the Persuasiveness of Language Models. Anthropic. https://www.anthropic.com/news/measuring-model-persuasiveness. Accessed: 2024-5-13.

Eckert, C. and Isaksson, O. 2017. Safety Margins and Design Margins: A Differentiation between Interconnected Concepts. *Procedia CIRP* 60: 267–272. doi.org/10.1016/j.procir.2017.03.140.

Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; DasSarma, N.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Jones, A.; Kernion, J.; Lovitt, L.; Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2021. *A Mathematical Framework for Transformer Circuits*. San Francisco, CA: Anthropic. https://transformer-circuits.pub/2021/framework/index.html.

Falk, A., and Heckman, J. J. 2009. Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science* 326(5952): 535–538. doi.org/10.1126/science.1168244.

Feffer, M.; Sinha, A.; Lipton, Z. C.; and Heidari, H. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? arXiv preprint. arXiv:2401.15897 [cs.CY]. Ithaca, NY: Cornell University Library.

Franke, E., and Alexander, B. 2019. The Potential Influence of Artificial Intelligence on Plagiarism: A Higher Education Perspective. In Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics, eds. Griffiths, P., and Kabir, M. N. Reading, UK: Academic Conferences and Publishing International Limited.

Frontier Model Forum. 2023. Issue Brief: What is Red Teaming? https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf. Accessed: 2024-5-13.

Future of Life Institute. 2023. Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Accessed: 2024-5-13.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, S.; Ndousse, K.; and Jones, A. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv preprint. arXiv:2209.07858v2 [cs.CL]. Ithaca, NY: Cornell University Library.

Ghorbani, B.; Firat, O.; Freitag, M.; Bapna, A.; Krikun, M.; Garcia, X.; Chelba, C.; and Cherry, C. 2021. Scaling Laws for Neural Machine Translation. arXiv preprint. arXiv:2109.07740 [cs.LG]. Ithaca, NY: Cornell University Library.

Goldstein, J.; Chao, J.; Grossman, S.; Stamos, A.; and Tomz, M. 2024. How Persuasive Is AI-generated Propaganda? *PNAS Nexus* 3(2): 1–7. https://doi.org/10.1093/pnasnexus/pgae034.

Goldstein, J., and Lohn, A. 2023. Finding Language Models in Influence Operations. Lawfare. https://www.lawfaremedia.org/article/finding-language-models-in-influence-operations. Accessed: 2024-04-29.

Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; and Evans, O. 2018. Viewpoint: When Will AI Exceed Human Performance? Advice from Experts. *Journal of Artificial Intelligence Research* 62: 729–754. doi.org/10.1613/jair.1.11222.

Greenblatt, Ryan, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. "Stress-Testing Capability Elicitation With Password-Locked Models." arXiv:2405.19550 [cs.LG]. Ithaca, NY: Cornell University Library.

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: Association for Computational Logistics. doi.10.18653/v1/2022.acl-long.234.

Hazell, J. 2023. Large Language Models Can Be Used to Effectively Scale Spear Phishing Campaigns. arXiv preprint. arXiv:2305.06972 [cs.CY]. Ithaca, NY: Cornell University Library.

Heim, L.; Fist, T.; Egan, J.; Huang, S.; Zekany, S.; Trager, R.; Osborne, M. A.; and Zilberman, N. 2024. Governing through the Cloud: The Intermediary Role of Compute Providers in AI Regulation. arXiv preprint. arXiv:2403.08501 [cs.CY]. Ithaca, NY: Cornell University Library.

Hicks, M.-L.; Guest, E.; Whittlestone, J.; Ohrvik-Stott, J.; Zakaria, S.; Politi, C.; Ang, C.; Wade, I.; and Gunashekar, S. 2023. *Exploring Red Teaming to Identify New and Emerging Risks from AI Foundation Models*. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/conf_proceedings/CFA3031-1.html.

Hoffman, M., and Frase, H. 2023. *Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework*. Washington, DC: Center for Security and Emerging Technology. doi.org/10.51593/20230022.

Jain, A.; Singh, R.; and Vatsa, M. 2018. On Detecting GANs and Retouching Based Synthetic Alterations. In 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems. New York, NY: Institute of Electrical and Electronics Engineers. doi.com/10.1109/BTAS.2018.8698545.

Janis, I. L. 1997. Groupthink. In *Leadership: Understanding the Dynamics Of Power And Influence in Organizations*, edited by, R.P. Vecchio. Notre Dame, IN: University of Notre Dame Press. (Reprinted from *Psychology Today*, Nov. 1971: 43, 44, 46, 74–76). https://agcommtheory.pbworks.com/f/GroupThink.pdf.

Ji, J. 2023. *What Does Red-Teaming Actually Mean?* Washington, DC: Center for Security and Emerging Technology. https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/. Accessed: 2024-04-29.

Kalla, J. L., and Broockman, D. E. 2018. The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments. *American Political Science Review* 112(1): 148–166. doi.org/:10.1017/S0003055417000363.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361v1 [cs.LG]. Ithaca, NY: Cornell University Library.

Kapoor, Sayash, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins et al. "On the Societal Impact of Open Foundation Models." arXiv:2403.07918 [cs.CY]. Ithaca, NY: Cornell University Library.

Khlaaf, H. 2023. *Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems*. New York: Trail of Bits.

King, A. 2024. King, Colleagues Unveil Bipartisan Framework to Identify, Minimize Artificial Intelligence Risks. Office of US Senator Angus King. https://www.king.senate.gov/newsroom/press-releases/king-colleagues-unveil-bipartisan-framework-to-identify-minimize-artificial-intelligence-risks. Accessed: 2024-04-29.

Kreps, S., and Kriner, D. L. 2023. The Potential Impact of Emerging Technologies on Democratic Representation: Evidence from a Field Experiment. *New Media & Society* 0(0). doi.org/10.1177/14614448231160526.

Kreps, S.; McCain, R. M.; and Brundage, M. 2022. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* 9(1): 104–117. doi:10.1017/XPS.2020.37.

Kuehn, A., and Mueller, M. 2014. Analyzing Bug Bounty Programs: An Institutional Perspective on the Economics of Software Vulnerabilities. Paper presented at the TPRC 42nd Research Conference on Communication, Information and Internet Policy. George Mason University School of Law. Arlington, Virginia, September 12–14. doi.org/10.2139/ssrn.2418812.

Levitt, S. D., and List, J. A. 2009. Field Experiments in Economics: The Past, the Present, and the Future. *European Economic Review* 53:1-18. doi.org/10.3386/w14356.

Lin, H. 2016. Attribution of Malicious Cyber Incidents: From Soup to Nuts. *Journal of International Affairs* 70(1): 75–137. https://www.jstor.org/stable/90012598.

Lin, Z.; Cui, J.; Liao, X.; and Wang, X. 2024. *Malla*: Demystifying Real-World Large Language Model Integrated Malicious Services. arXiv preprint. arXiv:2401.03315v1 [cs.CR]. Ithaca, NY: Cornell University Library.

Longpre, S.; Kapoor, S.; Klyman, K.; Ramaswami, A.; Bommasani, R.; Blili-Hamelin, B.; Huang, Y.; Skowron, A.; Yong, Z.; Kotha, S.; et al. 2024. A Safe Harbor for AI Evaluation and Red Teaming. arXiv preprint. arXiv:2403.04893 [cs.AI]. Ithaca, NY: Cornell University Library.

Malhotra N.; Monin, B.; and Tomz, M. 2019. Does Private Regulation Preempt Public Regulation? *American Political Science Review* 113(1): 19-37. doi:10.1017/S0003055418000679.

McGuffie, K., and Newhouse, A. 2020. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. arXiv preprint. arXiv:2009.06807 [cs.CY]. Ithaca, NY: Cornell University Library.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY: Association for Computing Machinery. doi.org/10.1145/3287560.3287596.

Mouton, C. A.; Lucas, C.; and Guest, E. 2024. *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*. Santa Monica, CA: RAND Corporation. doi.org/10.7249/RRA2977-2.

Mullins, C. 2018. *Retrospective Analysis of Long-Term Forecasts*. Alexandria, VA: Bryce Space and Technology. https://www.openphilanthropy.org/files/Blog/Mullins_Retrospective_Analysis_Longterm_Forecasts_Final_Report.pdf.

Musser, M. 2023. A Cost Analysis of Generative Language Models and Influence Operations. arXiv preprint. arXiv:2308.03740 [cs.CY]. Ithaca, NY: Cornell University Library.

OpenAI. 2023a. GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf. Accessed: 2024-5-13.

OpenAI. 2023b. OpenAI Red Teaming Network. https://openai.com/index/red-teaming-network/. Accessed: 2024-5-2.

Orth, T., and Bialik, C. 2023. Majorities of Americans Are Concerned about the Spread of AI Deepfakes and Propaganda. YouGov. https://today.yougov.com/technology/articles/46058-majorities-americans-are-concerned-about-spread-ai. Accessed: 2024-04-29.

Pastor-Galindo, J.; Nespoli, P.; Gómez Mármol, F.; and Pérez, G. M. 2020. The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends.

In *IEEE Access* 8: 10282–10304. doi.org/10.1109/ACCESS.2020.2965257.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. arXiv preprint. arXiv:2202.03286 [cs.CL]. Ithaca, NY: Cornell University Library.

Phuong, M.; Aitchison, M.; Catt, E.; Cogan, S.; Kaskasoli, K.; Krakovna, V.; Lindner, D.; Rahtz, R.; Assael, Y.; Hodkinson, S.; Howard, H.; Lieberum, T.; Kumar, R.; Abi Raad, M.; Webson, A.; Ho, L.; Lin, S.; Farquhar, S.; Hutter, M.; Delétang, G.; Ruoss, A.; El-Sayed, S.; Brown, S.; Dragan, A.; Shah, R.; Dafoe, A.; and Shevlane, T. 2024. Evaluating Frontier Models for Dangerous Capabilities. arXiv preprint. arXiv:2403.13793 [cs.LG]. Ithaca, NY: Cornell University Library.

Poldrack, R. A.; Lu, T.; and Beguš, G. 2023. AI-Assisted Coding: Experiments with GPT-4. arXiv preprint. arXiv: 2304.13187v1 [cs.AI]. Ithaca, NY: Cornell University Library.

Quillian, L., and Midtbøen, A. H. 2021. Comparative Perspectives on Racial Discrimination in Hiring: The Rise of Field Experiments. *Annual Review of Sociology* 47: 391–415. doi.org/10.1146/annurev-soc-090420-035144.

Rehman, I. 2019. Facebook-Cambridge Analytica Data Harvesting: What You Need to Know. *Library Philosophy and Practice* (e-journal): 2497. https://core.ac.uk/download/pdf/220153793.pdf.

Salvi, F.; Ribeiro, M. H.; Gallotti, R.; and West, R. 2024. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial. arXiv preprint. arXiv:2403.14380.

Saunders, W.; Yeh, C.; Wu, J.; Bills, S.; Ouyang, L.; Ward, J.; and Leike, J. 2022. Self-Critiquing Models for Assisting Human Evaluators. arXiv preprint. arXiv:2206.05802v2 [cs.CL]. Ithaca, NY: Cornell University Library.

Schumer, C. 2023. Majority Leader Schumer Delivers Remarks to Launch SAFE Innovation Framework for Artificial Intelligence at CSIS. Washington, DC: Senate Democrats. https://www.democrats.senate.gov/news/press-releases/majority-leader-schumer-delivers-remarks-to-launch-safe-innovation-framework-for-artificial-intelligence-at-csis. Accessed: 2024-04-29.

Scott, M.; Volpicelli, G.; Chatterjee, M.; Manancourt, V.; Goujard, C.; and Bordelon, B. 2024. Inside the Shadowy Global Battle to Tame the World's Most Dangerous Technology. POLITICO. https://www.politico.eu/article/ai-control-kamala-harris-nick-clegg-meta-big-tech-social-media/. Accessed: 2024-04-29.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; Ho, L.; Siddarth, D.; Avin, S.; Hawkins, W.; Kim, B.; Gabriel, I.; Bolina, V.; Clark, J.; Bengio, Y.; Christiano, P.; and Dafoe, A. 2023. Model Evaluation for Extreme Risks. arXiv preprint. arXiv:2305.15324 [cs.AI]. Ithaca, NY: Cornell University Library.

Soice, E. H.; Rocha, R.; Cordova, K.; Specter, M.; and Esvelt, K. M. 2023. Can Large Language Models Democratize Access to Dual-Use Biotechnology? arXiv preprint. arXiv:2306.03809 [cs.CY]. Ithaca, NY: Cornell University Library.

Solaiman, I. 2023. The Gradient of Generative AI Release: Methods and Considerations. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery. doi.org/10.1145/3593013.3593981.

Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; Daumé, H.; Dodge, J.; Evans, E.; Hooker, S.; Jernite, Y.; Luccioni, A. S.; Lusoli, A.; Mitchell, M.; Newman, J.; Png, M.-T.; Strait, A.; and Vassilev, A. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. arXiv preprint. arXiv:2306.05949 [cs.CY]. Ithaca, NY: Cornell University Library.

Spitale, G.; Biller-Andorno, N.; and Germani, F. 2023. AI Model GPT-3 (Dis)informs Us Better than Humans. *Science Advances* 9(26): eadh1850. doi.org/10.1126/sciadv.adh1850.

Steiger, M.; Bharucha, T. J.; Venkatagiri, S.; Riedl, M. J.; and Lease, M. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York: Association for Computing Machinery. doi.org/10.1145/3411764.3445092.

Steinhardt, J. 2022. AI Forecasting: One Year In. *Bounded Regret* (blog). https://bounded-regret.ghost.io/ai-forecasting-one-year-in/. Accessed: 2024-04-29.

Urbina, F., Lentzos, F., Invernizzi, C. et al. 2022. Dual Use of Artificial-Intelligence-Powered Drug Discovery. *Nature Machine Intelligence* 4: 189–191. doi.org/10.1038/s42256-022-00465-9.

Wald, H., and Shojania, K. 2001. Incident Reporting. In *Making Health Care Safer: A Critical Analysis of Patient Safety Practices.* Evidence Report/Technology Assessment 43. AHRQ Publication 01-E058. Rockville, MD: Agency for Healthcare Research and Quality. https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/.

Wallace, E.; Xiao, K.; Leike, R.; Weng, L.; Heidecke, J.; and Beutel, A. 2024. The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. arXiv preprint. arXiv:2404.13208v1 [cs.CR]. Ithaca, NY: Cornell University Library.

Walsh, E. 2014. How to Better Research the Possible Threats Posed by AI-Driven Misuse of Biology. Bulletin of the Atomic Scientists. https://thebulletin.org/2024/03/how-to-better-research-the-possible-threats-posed-by-ai-driven-misuse-of-biology/. Accessed: 2024-04-29.

Ward, C.; Polglase, K.; Shukla, S.; Mezzofiore, G.; and Lister, T. 2020. Russian Election Meddling is Back – via Ghana and Nigeria – and in Your Feeds. CNN. https://www.cnn.com/2020/03/12/world/russia-ghana-troll-farms-2020-ward/index.html. Accessed: 2024-04-29.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of Risks Posed by Language Models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery.

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. Defending against Neural Fake News. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, edited by H. Wallach et al. Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Inc.

Zellner, M.; Abbas, A. E.; Budescu, D. V.; and Galstyan, A. 2021. A Survey of Human Judgement and Quantitative Forecasting Methods. *Royal Society Open Science* 8(2): 1–29. doi.org/10.1098/rsos.201187.

Zenco, M. 2015. *Red Team: How to Succeed by Thinking Like the Enemy*. New York: Basic Books.

Zhang, B.; Dreksler, N.; Anderljung, M.; Kahn, L.; Giattino, C.; Dafoe, A.; and Horowitz, M. C. 2022. Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers. arXiv preprint. arXiv:2206.04132v1 [cs.CY]. Ithaca, NY: Cornell University Library.

Zhong, L.; Wang, Z.; and Shang, J. 2024. LDB: A Large Language Model Debugger via Verifying Runtime Execution Step by Step. arXiv preprint. arXiv:2402.16906v4 [cs.SE]. Ithaca, NY: Cornell University Library.