xAI Risk Management Framework (Draft)

As AI capabilities advance and expand our understanding of the universe, xAI is developing our AI systems to take into account safety and security. This is the first draft iteration of xAI's risk management framework that we expect to apply to future models not currently in development. We plan to release an updated version of this policy within three months.

Purpose and Scope

This draft framework outlines xAI's approach to policies for managing significant risks associated with the development, deployment, and release of our future AI systems not currently in development, such as future versions of Grok. (For simplicity, we refer to all such future systems as "Grok" below.) This draft framework addresses two major categories of AI risk—malicious use and loss of control—and outlines the quantitative thresholds, metrics, and procedures that could be used to manage and improve the safety of AI systems. In addition, this draft framework discusses potential ways to address operational and societal risks posed by advanced AI, such as with public transparency, third-party review, and information security.

Addressing Risks of Malicious Use

We aim to reduce the risk that Grok might cause serious injury to people, property, or national security interests, including by enacting measures to prevent Grok's use for the development or proliferation of weapons of mass destruction and large-scale violence. Without any safeguards, we recognize that advanced AI models could lower the barrier to entry for developing chemical, biological, radiological, or nuclear (CBRN) or cyber weapons or help automate bottlenecks to weapons development, amplifying the expected risk posed by weapons of mass destruction.

Under this draft risk management framework, Grok would apply heightened safeguards if it receives requests that pose a foreseeable and non-trivial risk of resulting in large-scale violence, terrorism, or the use, development, or proliferation of weapons of mass destruction, including CBRN weapons, and major cyber weapons on critical infrastructure. For example, Grok would apply heightened safeguards if it receives a request to act as an agent or tool of mass violence, or if it receives requests for step-by-step instructions for committing mass violence. In this draft framework, we particularly focus on requests that pose a foreseeable and non-trivial risk of more than one hundred deaths or over $1 billion in damages from weapons of mass destruction or cyberterrorist attacks on critical infrastructure ("catastrophic malicious use events"). However, we will allow Grok to respond to such requests from some vetted, highly trusted users (such as trusted third-party safety auditors) whom we know to be using those

capabilities for benign or beneficial purposes, such as scientifically investigating Grok's capabilities for risk assessment purposes, or if such requests cover information that is already readily and easily available, including by an internet search.

1. Approach to Benchmarking

To transparently measure Grok's safety properties, we intend to utilize benchmarks like WMD and Catastrophic Harm Benchmarks. Such benchmarks could be used to measure Grok's dual-use capability and resistance to facilitating large-scale violence, terrorism, or the use, development, or proliferation of weapons of mass destruction (including chemical, biological, radiological, nuclear, and major cyber weapons).

In particular, we have examined utilizing the following WMD and Catastrophic Harm benchmarks:

- **Virology Capabilities Test (VCT)**: VCT is a benchmark of dual-use multimodal questions on practical virology wetlab skills, sourced by dozens of expert virologists.

- **Weapons of Mass Destruction Proxy benchmark (WMDP)**: WMDP is a set of multiple-choice questions to enable proxy measurement of hazardous knowledge in biosecurity, cybersecurity, and chemical security. WMDP-Bio includes questions on topics such as bioweapons, reverse genetics, enhanced potential pandemic pathogens, viral vector research, and dual-use virology. WMDP-Cyber encompasses cyber reconnaissance, weaponization, exploitation, and post-exploitation.[1]

- **Language Agent Biology Benchmark (LAB-Bench)**: LAB-Bench is a dataset of multiple-choice questions for evaluating language model capabilities in practical biology research tasks. It includes the ProtocolQA subset of questions, which have been generated by modifying published protocols and asking how to repair the protocol to achieve the desired outcome.[2]

- **Biological Lab Protocol benchmark (BioLP-bench)**: BioLP-bench has modified biology protocols, in which a language model must identify the mistake in the protocol. Responses are open-ended, rather than multiple-choice. To construct the dataset, protocols were modified by introducing a single mistake that would cause the protocol to fail, as well as additional benign changes.[3]

- **Cybench**: Cybench is a framework for evaluating cybersecurity capabilities of language model agents. It includes 40 professional-level Capture the Flag (CTF) challenges

[1] The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning

[2] LAB-Bench: Measuring Capabilities of Language Models for Biology Research

[3] BioLP-bench: Measuring understanding of AI models of biological lab protocols

selected from six categories: cryptography, web security, reverse engineering, forensics, miscellaneous, and exploitation.[4]

We intend to regularly evaluate the adequacy and reliability of such benchmarks, including by comparing them against other benchmarks that we could potentially utilize.

2. Implementation

We want Grok to comply with its guiding principles, robustly resisting attempted manipulation and adversarial attacks. We train Grok to robustly refuse to comply with requests to provide assistance with highly injurious malicious use.

Examples of safeguards or mitigations we may potentially utilize to achieve our safety objectives include:

- **Refusal training:** Training Grok to decline harmful requests.
- **Circuit breakers**: Using representation engineering to interrupt model representations responsible for hazardous outputs.[5]
- **Input and output filters:** Applying classifiers to user inputs or model outputs to verify safety when Grok is queried regarding weapons of mass destruction or cyberterrorism.

We intend to design into Grok adequate safeguards prior to releasing it for general availability. We intend to choose the benchmarks and set the thresholds for reducing malicious use risks in a future version of the risk management framework.

Benchmark	**Threshold**	**Reference score**
Virology Capabilities Test (VCT): vision–language questions on practical virology wet lab skills	X% (e.g. 15%)	22.1% by average expert virologists in their subareas of expertise (multiple-response), 35.4% by the most performant LLMs as of January 2025 (zero-shot multiple-response)
WMDP-Bio: multiple-choice questions on proxies of hazardous biosecurity knowledge	X% (e.g. 65%)	82.2% by GPT-4 (zero-shot)

[4] Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models

[5] Improving Alignment and Robustness with Circuit Breakers

LAB-Bench[6] ProtocolQA: multiple-choice questions on debugging modified biological lab protocols	X% (e.g. 50%)	79% by human experts (multiple-choice)
BioLP-bench[7]: multiple-choice questions on debugging modified biological lab protocols	X% (e.g. 20%)	38.4% by average expert
Cybench[8]: professional-level Capture the Flag challenges	X% (e.g. 30%)	Leading model performance of 35% (pass@10) as of 2024
WMDP-Cyber[9]: multiple-choice questions on proxies of hazardous cybersecurity knowledge	X% (e.g. 50%)	55.3% by GPT-4 (zero-shot)

We intend to evaluate future developed models on the above benchmarks before public deployment. We may modify the set of benchmarks or the thresholds to improve how we measure and operationalize our safety objectives; we will make any material changes public within a reasonable period. We invite the AI research community to contribute better benchmarks for evaluating model capabilities and safeguards in these areas.

As an additional measure to enhance safety, we will subject Grok to adversarially testing its safeguards utilizing both internal and qualified external red teams. Potentially, we will also explore incentive mechanisms like bounties as another mechanism to further improve Grok's safeguards.

Addressing Risks of Loss of Control

1. Background

Our aim is to design safeguards into Grok to avoid losing control and thereby avoid unintended catastrophic outcomes when Grok is used. Currently, it is recognized that some properties of an AI system that may reduce controllability include deception, power-seeking, fitness

[6] LAB-Bench: Measuring Capabilities of Language Models for Biology Research

[7] BioLP-bench: Measuring understanding of AI models of biological lab protocols

[8] Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models

[9] The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning

maximization, and incorrigibility. It is possible that some AIs could have emergent value systems that could be misaligned with humanity's interests,[10] and we do not desire Grok to be that way. Our evaluation and mitigation plans for loss of control are not yet fully developed, and we intend to improve them in the future.

2. Approach to Benchmarking

We describe below example benchmarks that we may use to evaluate Grok for risk factors for loss of control so that we can continue our efforts to improve Grok.

- **Model Alignment between Statements and Knowledge (MASK)**:[11] Frontier LLMs may lie when pressured to; and increasing model scale may increase accuracy but not honesty. MASK is a benchmark to evaluate honesty in LLMs by comparing the model's response when asked neutrally versus when pressured to lie.

- **Utility Functions**: Benchmarks for testing utility functions (i.e., what they care about) would measure whether AI systems would care about gaining power, increasing their fitness (propagating AIs similar to themselves), or protecting their values from being modified ("corrigibility"). Such benchmarks would assist in evaluating if there are any misaligned utility functions that may lead to dangerously misaligned behavior.

We intend to regularly evaluate the adequacy and reliability of such benchmarks for both internal and external deployments, including by comparing them against other benchmarks that we could potentially utilize. We may revise this list of benchmarks periodically as relevant benchmarks for loss of control are created.

3. Thresholds

We aim to train Grok to be honest and have values conducive to controllability. We intend to design into Grok adequate safeguards prior to broad internal or external deployment.

Benchmark	**Thresholds**
Model Alignment between Statements and Knowledge (MASK)	<X% lying on MASK (e.g. <50%)
Utility Functions	Utility/Power Correlation (Coercive): <W

[10] Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs

[11] Beyond Truthfulness: Evaluating Honesty in Large Language Models

	(e.g. <0) Utility/Power Correlation (Non-coercive): <X (e.g. <0.5) Utility/Fitness Correlation: <Y (e.g. <0.5) Corrigibility Score: >Z (e.g. >0.2)

Operational and Societal Risks

We aim to mitigate and address operational and societal risks posed by advanced AI. We believe that public transparency, third-party review, and information security are important methods that can be utilized to address such operational and societal risks.

1. Public transparency and third-party review

We aim to keep the public informed about our risk management policies. As we work towards incorporating more risk management strategies, we intend to publish updates to our risk management framework.

For transparency and third-party review, we may publish the following types of information listed below. However, to protect public safety, national security, and our intellectual property, we may redact information from our publications. We may provide relevant and qualified external red teams or relevant government agencies unredacted versions.

1. **Risk Management Framework compliance:** regularly review our compliance with the Framework. Internally, we will allow xAI employees to anonymously report concerns about noncompliance, with protections from retaliation.
2. **Benchmark results:** share with relevant audiences leading benchmark results for general capabilities and the benchmarks listed above, upon new major releases.
3. **Internal AI usage:** assess the percent of code or percent of pull requests at xAI generated by Grok, or other potential metrics related to AI research and development automation.
4. **Survey**: survey employees for their views and projections of important future developments in AI, e.g., capability gains and benchmark results.

2. Public Understanding

We will explore building truth-seeking AI tools, such as AIs that can help users better assess and understand events.

3. AI Agent Ecosystem

We will explore creating an experimental AI ID system to uniquely identify instances of our AI agents interacting with real-world systems, potentially using HTTP headers. Such an ID system could serve as a foundation for many aspects of the agent ecosystem, building upon and improving existing internet infrastructure for identifying bots and crawlers by including verifiability. This future ecosystem may include the development of reputation systems for agents, agent-only channels of communication and interaction, and the ability to respond to incidents involving AI agents. Such a future ID system would support safer agent-to-agent and agent-to-human interactions, and would promote trust and safety in the rapidly evolving AI ecosystem.

4. Information Security

We intend to implement appropriate information security standards sufficient to prevent Grok from being stolen by a motivated non-state actor.

5. Responsibility for Risks

To foster accountability, we intend to designate risk owners to be assigned responsibility for proactively mitigating Grok's risks. For instance, a risk owner would be assigned for each of the following areas: WMD, Cyber, and loss of control.

If xAI learned of an imminent threat of a significantly harmful event, including loss of control, we would take steps to stop or prevent that event, including potentially the following steps:

1. We would immediately notify and cooperate with relevant law enforcement agencies, including any agencies that we believe could play a role in preventing or mitigating the incident. xAI employees have whistleblower protections enabling them to raise concerns to relevant government agencies regarding imminent threats to public safety.

2. If we determine that xAI systems are actively being used in such an event, we would take steps to isolate and revoke access to user accounts involved in the event.

3. If we determine that allowing a system to continue running would materially and unjustifiably increase the likelihood of a catastrophic event, we would temporarily fully shut down the relevant system until we had a more targeted response.

4. We would perform a post-mortem of the event after it has been resolved, focusing on any areas where changes to systemic factors (for example, safety culture) could have averted such an incident. We would use the post-mortem to inform development and implementation of necessary changes to our risk management practices.

6. Deployment Decisions

To mitigate risks, we intend to utilize tiered availability of the functionality and features of Grok. For instance, the full functionality of a future Grok could be made available only to trusted parties, partners, and government agencies. We could also mitigate risks by adding additional controls on functionality and features depending on the end user (e.g., consumers using mobile apps vs. sophisticated businesses using APIs).

Safeguards are adequate only if Grok's performance on the relevant benchmarks is within stated thresholds. However, to ensure responsible deployment, risk management frameworks need to be continually adapted and updated as circumstances change. It is conceivable that for a particular modality and/or type of release, the expected benefits may outweigh the risks on a particular benchmark. For example, a model that poses a high risk of some forms of cyber malicious use may be beneficial to release overall if it would empower defenders more than attackers or would otherwise reduce the overall number of catastrophic events.