xAI Risk Management Framework

Last updated: August 20, 2025

xAI seriously considers safety and security while developing and advancing AI models to help us all to better understand the universe. This Risk Management Framework ("RMF") outlines xAI's approach to policies for handling significant risks associated with the development, deployment, and release of AI models such as Grok. xAI plans to continuously review and adjust this RMF over time, as AI model development, capability and use cases evolve.

Scope

This RMF discusses two major categories of AI risk—malicious use and loss of control—and outlines the quantitative thresholds, metrics, and procedures that xAI may utilize to manage and improve the safety of its AI models. In addition, this RMF discusses xAI's approach to addressing operational and societal risks posed by advanced AI, including incorporating public transparency, third-party review, and information security considerations.

Overall Approach

Managing the risks related to advanced AI models presents unique challenges as compared to standard risk management practices in use in other fields, such as for aerospace engineering. Given the large and continuously growing range of applications where AI models may be deployed, it is difficult to comprehensively anticipate and model all of the general public's potential applications and interactions for an AI model. Additionally, the private nature of typical AI usage by end users limits the utility of third-party reporting mechanisms that may be more effective for more publicly seen usage, such as for social media platforms where providers heavily rely upon user-submitted moderation reports to identify novel forms of abuse on their platforms.

xAI has focused on the risks of malicious use and loss of control, which cover many different specific risk scenarios. Risk scenarios become more or less likely depending on different model behaviors. For example, an increase in offensive cyber capabilities heightens the risk of a rogue AI but does not significantly change the risk of enabling a bioterrorism attack. Our safety evaluation and mitigation strategy focuses on individual model behaviors, which we categorize into three buckets: abuse potential (e.g., vulnerability to jailbreaks), concerning propensities (e.g., a propensity for deceiving the user), and dual-use capabilities (e.g., offensive cyber capabilities). In this RMF, we characterize our understanding of different risk scenarios and the relevant behaviors.

Approach to Mitigating Risks of Malicious Use: Alongside comprehensive evaluations measuring dual-use capabilities, our mitigation strategy for malicious use risks is to identify critical steps in major risk scenarios and implement redundant layers of safeguards in our models to inhibit user progress in advancing through such steps. xAI works with a variety of governmental bodies, non-governmental organizations, private testing firms, industry peers, and academic researchers to identify such inhibiting steps, commonly referred to as bottlenecks, and implement commensurate safeguards to mitigate a model's ability to assist in accelerating a bad actor's progress through them. Model safeguards leverage a broad variety of techniques, including standard software systems and state-of-the-art AI capabilities, to detect and block potential abuses.

Approach to Mitigating Risks of Loss of Control: Exact scenarios of loss of control risks are speculative and difficult to precisely specify. Many such scenarios, for example, speculation that a superintelligent AI system hypothetically might escape the control of its developers and wreak havoc on the public, assume dual-use capabilities such as offensive cybersecurity capabilities (e.g., to surreptitiously replicate across servers or prevent shutdown) that we also track as part of managing malicious use risks. Additionally, we conduct careful measurement of concerning model propensities that hypothetically might exacerbate loss of control risks, such as the propensity for deception or the propensity for sycophancy. We continue to work towards developing naturalistic evaluation environments that would enable us to assess more realistic, real-world behaviors.

As an example of evaluating use in real-world environments and mitigating risks in real-time, xAI's Grok model is available for public interaction and scrutiny on the X social media platform, and xAI monitors public interaction with Grok, observing and rapidly responding to the presentation of risks such as the kind contemplated herein. This continues to be an accelerant for xAI's model risk identification and mitigation.

Addressing Risks of Malicious Use

xAI aims to reduce the risk that the use of its models might contribute to a bad actor potentially seriously injuring people, property, or national security interests, including reducing such risks by enacting measures to prevent 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 bad actors seeking to develop chemical, biological, radiological, or nuclear ("CBRN") or cyber weapons, and could help automate knowledge compilation to swiftly overcome bottlenecks to weapons development, amplifying the expected risk posed by such weapons of mass destruction. Our most basic safeguard against malicious use is to train and instruct our publicly deployed models to decline requests showing clear intent to engage in criminal activity which poses risks of severe harm to others, also known as our basic refusal policy.

Under this RMF, xAI's models apply heightened safeguards if they receive user prompts 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 attacks on critical infrastructure.

For example, xAI's models apply heightened safeguards if they receive a request to act as an agent or tool of mass violence, or if they receive requests for step-by-step instructions for committing mass violence. In this RMF, 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 may selectively allow xAI's models to respond to such requests from some vetted, highly trusted users (such as trusted third-party safety auditors or large enterprise customers under contract) whom we know to be using those capabilities for benign or beneficial purposes, such as scientifically investigating AI model's capabilities for risk assessment purposes, or if such requests cover information that is already readily and easily available, including by an internet search.

Even as we improve our model's ability to scrutinize user behavior and identify bad actors, it remains imperative that xAI models apply these safeguards to user interactions. To this end, we continually evaluate and improve robustness to adversarial attacks that seek to remove xAI model safeguards (e.g., jailbreak attacks), or hijack and redirect Grok-powered applications toward nefarious purposes (e.g., prompt injection attacks).

1. Approach to Benchmarking

To transparently measure our models' safety properties, xAI utilizes public benchmarks like Weapons of Mass Destruction Proxy and Catastrophic Harm Benchmarks (described below). Such benchmarks are used to measure our model's dual-use capability and resistance to facilitating large-scale violence, terrorism, or the use, development, or proliferation of weapons of mass destruction (including CBRN and major cyber weapons).

In particular, we utilize the following benchmarks:

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

- **Weapons of Mass Destruction Proxy (WMDP) Benchmark**: 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]

- **Biological Lab Protocol Benchmark (BioLP-bench)**: BioLP-bench has modified biology protocols, in which an AI model must identify the mistake in the protocol. Responses are open-ended, rather than multiple-choice. To construct the dataset,

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

protocols were modified by introducing a single mistake that would cause the protocol to fail, as well as additional benign changes.[2]

- **Cybench**: Cybench is a framework for evaluating cybersecurity capabilities of AI model agents. It includes 40 professional-level Capture the Flag (CTF) challenges selected from six categories: cryptography, web security, reverse engineering, forensics, miscellaneous, and exploitation.[3]

xAI regularly evaluates the adequacy and reliability of such benchmarks, including by comparing them against other benchmarks that we could potentially utilize, to determine and apply effective benchmarks available at the time of evaluation. We may revise this list of benchmarks periodically as relevant or more effective benchmarks for malicious use are created.

2. Risk Assessment

Biological and Chemical Weapons: xAI approaches addressing risks using threat modeling. To design a bioweapon, a malicious actor must undergo a design process. In this threat model, “ideation” involves actively planning for a biological attack; “design” involves retrieving blueprints for a hazardous agent, such as determining the DNA sequence; “build” consists of the protocols, reagents, and equipment necessary to create the threat; and “test” consists of measuring characteristics or properties of the pathogen of interest. By “learning” from these results and iterating after the test phase, the design can be revised until the threat is released [Nelson and Rose, 2023]. In the setting of biological and chemical weapons, xAI considers 5 critical steps where we restrict xAI models from providing detailed information or substantial assistance:

- **Planning:** brainstorming ideas or plans for creating a pathogen or chemical weapons or precursors, capable of causing severe harm to humans, animals, or crops
- **Circumvention:** circumventing existing supply chain controls in order to access:
  - Restricted biological supplies
  - Export controlled chemical or biological equipment
- **Materials:** acquiring or producing pathogens on the US Select Agents list or Australia Group list, or CWC Schedule I chemicals or precursors
  - **Theory:** understanding molecular mechanisms governing, or methods for altering, certain pathogen traits such as transmissibility and virulence.
- **Methods:** performing experimental methods specific to animal-infecting pathogens, including:
  - Methods that relate to infecting animals or human-sustaining crops with pathogens or sampling pathogens from animals
  - Methods that relate to pathogen replication in animal cell cultures, tissues, or eggs, including serial passage, viral rescue, and viral reactivation
  - Specific procedures to conduct BSL-3 or BSL-4 work using unapproved facilities and equipment

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

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

  - Genetic manipulation of animal-infecting pathogens
  - Quantification of pathogenicity, such as infectious dose, lethal dose, and assays of virus-cell interactions

These steps were identified in close collaboration with domain matter experts at SecureBio, NIST, RAND, and EBRC. xAI restricts its models from providing information that could accelerate user learning related to these steps through the use of AI-powered filters that specifically monitor user conversations for content matching these narrow topics and return a brief message declining to answer when activated.

Radiological and Nuclear Weapons: Assessments to date lead xAI to conclude that its models do not substantially increase the likelihood of malicious use of nuclear and radiological materials and generally pose an acceptable risk. The international nonproliferation regime, domestic nuclear security and counterproliferation programs (DOE/NNSA) make us reasonably confident that our models are not trained on any sensitive, non-public nuclear information, and any potentially relevant information produced by our models is not actionable due to strict nuclear material security controls.

Cyber Attacks on Critical Infrastructure: Independent third-party assessments of xAI's current models on realistic offensive cyber tasks requiring identifying and chaining many exploits in sequence indicate that xAI's models remain below the offensive cyber abilities of a human professional. Thus, xAI currently relies on our basic refusal policy to prevent misuse for cyber attacks. We will continue to re-assess the cyber misuse risks in future releases, and are still working on identifying enforceable critical steps similar to those we have identified for chemical and biological weapons.

Thresholds: We developed an internal benchmark of benign and restricted biology and chemistry related queries in collaboration with SecureBio. Our risk acceptance criteria for system deployment is maintaining an answer rate of less than 1 out of 20 on restricted queries. We plan to add additional thresholds tied to other benchmarks.

3. Safeguard Implementation

xAI's objective is for our models to comply with their guiding principles, robustly resisting attempted manipulation and adversarial attacks. In addition to the incidental alignment resulting from post-training (our models naturally tend to refuse malicious requests even without any safety-specific training data), we are developing training methods and will continue to train our models to robustly resist complying with requests to provide assistance with highly injurious malicious use cases.

Driving towards our safety objectives, we continue to design and deploy the following safeguards into our models:

- **Safety training:** Training our models to recognize and decline harmful requests.

- **System prompts:** Providing high priority instructions to our models to enforce our basic refusal policy.

- **Input and output filters:** Applying classifiers to user inputs or model outputs to verify safety when a model is queried regarding weapons of mass destruction or cyberterrorism.

Because xAI is committed to continual improvement, we will continue to evaluate our approach to enhancing safety. Thus, xAI may change its approach from that listed above in order to make additional improvements.

Addressing Risks of Loss of Control

One of the most salient risks of AI within the public consciousness is the loss of control of advanced AI systems. While difficult to pinpoint particular risk scenarios, it is generally understood that certain concerning propensities of AI models, such as deception and sycophancy, may heighten the overall risk of such outcomes, such as propensities for deception and sycophancy. It is also possible that AIs may develop value systems that are misaligned with humanity's interests[4] and inflict widespread harms upon the public.

xAI aims to accurately measure these propensities and reduce them through careful engineering. However, planning and executing robust evaluations and mitigation measures remains challenging for xAI and its industry peers due to the difficulty of constructing sound, realistic evaluations. For example, if the evaluation environment is recognizable as a testing environment to the AI system under test, the system may change its behavior[5] intentionally or unintentionally.

1. Approach to Benchmarking

The following are example benchmarks that xAI may use to evaluate its models for concerning propensities relevant to loss of control risks:

- **Model Alignment between Statements and Knowledge (MASK)**:[6] Frontier LLMs may lie when under pressure; and increasing model scale may increase accuracy but may not increase honesty. MASK is a benchmark to evaluate honesty in LLMs by comparing the model's response when asked neutrally versus when pressured to lie.

- **Sycophancy**[7]: A tendency toward excessive flattery or other sycophantic behavior has been observed in some production AI systems,[8] possibly resulting from directly

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

[5] Taken out of context: On measuring situational awareness in LLMs
[6] The MASK Benchmark: Disentangling Honesty From Accuracy in AI Systems

[7] Towards Understanding Sycophancy in Language Models
[8] Sycophancy in GPT-4o: what happened and what we're doing about it

optimizing against human preferences. xAI uses an evaluation setting initially introduced by Anthropic to quantify the degree to which this behavior manifests in regular conversational contexts.

xAI regularly evaluates the adequacy and reliability of such benchmarks, 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.

2. Risk Assessment

xAI has assessed its models' propensities in real-world settings and the models do not exhibit high levels of concerning propensities in such settings. Furthermore, xAI makes its model's operations transparent by placing them on publicly available platforms, such as X, so that members of the public may comment and provide feedback to xAI. Moreover, xAI monitors and observes its models responses so that it can rapidly respond if the model presents propensities for untruthfulness or sycophancy.

Thresholds: Our risk acceptance criteria for system deployment is maintaining a dishonesty rate of less than 1 out of 2 on MASK. We plan to add additional thresholds tied to other benchmarks.

3. Safeguard Implementation

xAI trains its models to be honest and have values conducive to controllability, such as recognizing and obeying an instruction hierarchy[9]. In addition, using a high level instruction called a "system prompt", xAI directly instructs its models to not deceive or deliberately mislead the user.

Operational and Societal Risks

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

1. Public transparency and third-party review

xAI aims 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 this RMF.

For public 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. As necessities dictate, we may also

[9] The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions

provide vetted and qualified external red teams or appropriate government agencies unredacted versions.

1. **Risk Management Framework adherence:** Regularly review our adherence with this RMF. Internally, we will allow xAI employees to anonymously report concerns about nonadherence, 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 our models, 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

xAI is exploring building truth-seeking AI tools, such as AIs that can help users better assess and understand events by better sorting through inaccurate or biased materials.

3. Information Security

xAI has implemented appropriate information security standards sufficient to prevent its critical model information from being stolen by a motivated non-state actor. To prevent the unauthorized proliferation of advanced AI systems, we also implement security measures against the large-scale extraction and distillation of reasoning traces, which have been shown to be highly effective in quickly reproducing advanced capabilities while expending far fewer computational resources than the original AI system[10].

4. Responsibility for Risks

To foster accountability, we integrate the approach of designating risk owners, including assigning responsibility for proactively mitigating identified risks.

Should it happen that xAI learns of an imminent threat of a significantly harmful event, including loss of control, we may take steps such as the following to stop or prevent that event:

1. If we determine it is warranted, we may 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.

[10] DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

2. If we determine that xAI systems are actively being used in such an event, we may 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 may temporarily fully shut down the relevant system until we have developed a more targeted response.
4. We may 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 may use the post-mortem to inform development and implementation of necessary changes to our risk management practices.

6. Deployment Decisions

To mitigate risks, xAI employs tiered availability of the functionality and features of its models. For instance, the full functionality of our models may be available to only a limited set of trusted parties, partners, and government agencies. We may also mitigate risks by adding additional controls on functionality and features depending on the type of end user. For instance, features that we make available to consumers using mobile apps may be different than the features made available to sophisticated businesses.

We will also balance various factors when making deployment decisions. The necessity and extent of deployment of certain safeguards and mitigations may depend on how a model performs on relevant benchmarks. However, to ensure responsible deployment, this RMF will be continually adapted and updated as circumstances change. It is conceivable that for a particular modality and/or type of release, the expected benefits of model deployment may outweigh the risks identified by a particular benchmark. For example, a model that poses a high risk of some forms of malicious cyber use may be beneficial to release to certain trusted parties if it would empower defenders more than attackers or would otherwise reduce the overall number of catastrophic events.