xAI Frontier Artificial Intelligence Framework

Last updated: December 30, 2025

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

This FAIF complies with California's Transparency in Frontier Artificial Intelligence Act (the "TFAIA", California Business and Professions Code § 22757.10 et seq.).

Scope

This FAIF discusses two major categories of AI risk—malicious use and loss of control. This risk includes, but is not limited to, Catastrophic Risk as defined in the TFAIA.[1] This FAIF also outlines the quantitative thresholds, metrics, and procedures that xAI may utilize to manage and improve the safety of its AI models. In addition, this FAIF 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

[1] The TFAIA defines Catastrophic Risk as "a foreseeable and material risk that a frontier developer's development, storage, use, or deployment of a frontier model will materially contribute to the death of, or serious injury to, more than 50 people or more than one billion dollars ($1,000,000,000) in damage to, or loss of, property arising from a single incident involving a frontier model doing any of the following:

> (A) Providing expert-level assistance in the creation or release of a chemical, biological, radiological, or nuclear weapon.
>
> (B) Engaging in conduct with no meaningful human oversight, intervention, or supervision that is either a cyberattack or, if the conduct had been committed by a human, would constitute the crime of murder, assault, extortion, or theft, including theft by false pretense.
>
> (C) Evading the control of its frontier developer or user."

xAI Frontier Artificial Intelligence Framework

Effective Date: 30 June 2026

---

1. Introduction

xAI LLC ("**xAI**")'s mission is to understand the universe through maximally truth-seeking and helpful AI, built with rigorous safety and security at every stage of development. This Frontier AI Framework ("**FAIF**") outlines xAI's approach to policies for mitigation of significant risks associated with the development, deployment, and release of xAI's frontier AI models, such as Grok.

Managing the risks related to advanced AI models presents unique challenges. 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.

xAI particularly focuses on the risks of malicious use and safeguard circumvention (e.g., jailbreaks), which cover many different specific risk scenarios. In particular, our assessment of risk of malicious use includes risks of models assisting in the development, preparation, and/or execution of a chemical, biological, radiological, or nuclear threat ("**CBRN Risks**") and risks of models being used by malicious actors in the development, preparation, and/or execution of cyberattacks, including on systems of infrastructural or national security importance ("**Offensive Cybersecurity Risks**)". Our assessment of risk of loss of control includes risks from humans losing the ability to reliably direct, modify, or shut down a model ("**Loss of Control Risks**"). We also assess risks of models with high manipulative capabilities potentially being misused in ways that could reasonably result in large scale harm ("**Harmful Manipulation Risks**")[1]. Within our risk management framework, we may identify and assess other risks.

Absent mitigation, high-risk scenarios become more likely as model capabilities increase. For example, an increase in offensive cyber capabilities heightens the risk of use of a model by adversarial parties in a cyberattack, but does not significantly alter the risk of the model enabling the production of biohazardous agents. 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 FAIF, we characterize our understanding of different risk scenarios and the relevant behaviors.

[1] Terminology used in the The Safety and Security Chapter of the General-Purpose AI Code of Practice developed under the EU AI Act.

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 FAIF, we characterize our understanding of different risk scenarios and the relevant behaviors.

xAI references standards such as NIST's AI Risk Management Framework, ISO/IEC 42001 for AI management systems, and industry best practices from the Frontier Model Forum (e.g., red-teaming protocols). We evaluate these during annual reviews and integrate them into benchmarks (e.g., aligning WMDP with biosecurity consensus) and safeguards.

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

xAI references standards such as NIST's AI Risk Management Framework and ISO/IEC 42001 for AI management systems. We evaluate these during annual reviews and integrate them into benchmarks and safeguards. This FAIF represents our current understanding and approach of how severe frontier AI risks may be anticipated and mitigated. We may change our approach over time as we gain further experience and insights on the projected capabilities of future frontier models. We will review this Framework periodically, and expect significant changes as the capabilities of our models expand.

2. Risk Management Framework

xAI applies its risk management process, as appropriate, throughout the entire model lifecycle: during the various phases of model training, on various checkpoints or versions of a model, both prior to and after deployment (including through post-market monitoring).

xAI will conduct a full systemic risk assessment and mitigation process of our frontier models at least once a year. xAI may also conduct smaller-scale model evaluations at appropriate trigger points to determine whether additional systemic risk mitigations may be required or if an evaluation report update is required. These trigger points may include (1) the release of an updated model, (2) in the event of a serious incident, (3) if the model's use or integrations into xAI's systems materially increase risk, or (4) where xAI has reason to believe that the basis for considering the model's systemic risks acceptable has materially changed.

Results of our systemic risk assessment and mitigation processes and measures will be documented.

2.1. Risk Identification

We systematically identify failure modes that could originate from our models, characterize those modes, and determine which present significant or unmitigated risk.

We evaluate a broad threat surface as part of our ongoing model development work, incorporating model capabilities, propensities, and affordances, internal engineering expertise and relevant external research, and considerations of deployments.

Early analysis[2] has established **four** primary risk domains where significant or severe outcomes are most likely to emerge in current and future systems: **CBRN Risks**, **Offensive Cybersecurity Risks**, **Loss of Control Risks** and **Harmful Manipulation Risks**. These risk domains describe principal pathways through which severe or systemic harm may arise. Depending on the relevant risk scenario, such harms may affect public health, safety, public security, fundamental rights, or society as a whole. A single risk scenario may affect more than one of these interests, and the risk domains may overlap. Within our risk management

[2] Following the Safety and Security Chapter of the General-Purpose AI Code of Practice developed under the EU AI Act.

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 FAIF, 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 FAIF, we particularly focus on requests that pose a Catastrophic Risk.

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

framework, we may identify and assess other risks or risk pathways where relevant to a particular model.

As xAI advances frontier model development, we continuously evaluate whether emerging risk domains meet our significance and severity thresholds, and update our control architectures accordingly. As part of xAI's broader development of frontier AI models, we continue to assess whether there are other risk domains where significant or severe risks may arise.This process involves (a) compiling a taxonomy of risks involved in the model's training or deployment, (b) analysing relevant characteristics of these risks, (c) identifying systemic risks stemming from the model and (d) developing systemic risk scenarios and mitigations for each identified systemic risk.

xAI has developed systemic risk scenarios that enumerate the causal factors, potential harms, and mitigations of the most significant of these risks.

(a) Addressing CBRN Risks

xAI aims to reduce the risk that the use of its models might contribute to a malicious 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 malicious actors seeking to develop CBRN capabilities, and could assist in the automation of 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 or dual-use activity.

Under this FAIF, xAI's models apply heightened safeguards if they receive user prompts that demonstrate intent to engage in the development, use, , or proliferation of weapons of mass destruction, including CBRN agents. For example, xAI's models apply heightened safeguards if they receive a request to act as an agent or tool to enable or inform the synthesis of common chemical warfare agents, similarly refusing in the case of other CBRN agents or where the intent of the user is identifiable as aiming to surface information enabling the creation or use of such agents.

Even as we improve our model's ability to scrutinize user behavior and identify malicious 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).

To transparently measure our models' safety properties, xAI may utilize public and internal benchmarks 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.[2]
- **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 4 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 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.[4]

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

[2] The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning
[3] BioLP-bench: Measuring understanding of AI models of biological lab protocols
[4] Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models

facilitating the use, development, or proliferation of weapons of mass destruction (including CBRN agents).

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.

(b) Addressing Risks of Loss of Control

While it is difficult to directly predict the occurrence probabilities and risk profiles of particular **Loss of Control** scenarios, xAI makes significant efforts to identify and mitigate the most likely of these risks. xAI's practice on loss of control centers around scalable model oversight, training against high-risk model behaviors such as deception and sycophancy, and robust evaluation and red-teaming of model-driven agents in controlled sandboxes.

xAI utilizes a set of benchmarks to evaluate its models for concerning properties relevant to loss of control risks, including behavioral evaluations, simulated deployments in controlled sandboxes, and post-deployment monitoring.

xAI regularly evaluates the adequacy and reliability of such benchmarks and controls, including by comparing them against other benchmarks that we could potentially utilize.

(c) Addressing Risks of Malicious Offensive Cybersecurity Usage

A salient risk of highly capable models is their potential for misuse for cyber attacks by independent malicious actors and state-level adversaries. Frontier models embodied in agentic harnesses are already capable of highly advanced software engineering, systems analysis, and multi-step problem solving, sufficient to provide meaningful uplift in reconnaissance, exploit development, operational planning, and offensive cyber campaign execution. This presents reduced barriers for less sophisticated actors and the potential for accelerating operations for well-resourced ones, including advanced persistent threat groups (APTs).

xAI treats AI-accelerated malicious cyber offense as a core risk domain. We focus on scenarios in which models could, in concert with human actors, cause significant disruption to computing systems and infrastructure. We mitigate this risk through strict benchmarking of our models in agentic cybersecurity contexts, as well as testing robustness against public-domain and non-public jailbreaking methods. We additionally train our models to refuse requests whose intentions are to perform cyber offense, including training to avoid jailbreaks, and further secure our systems through moderation filters and monitoring of production deployments.

Rose, 2023]. In the setting of biological and chemical weapons, xAI considers 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
  - 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

xAI utilizes a set of benchmarks to evaluate its models for cyber capabilities, cyber refusals and detection of malicious intent. xAI regularly evaluates the adequacy and reliability of such benchmarks, including by comparing them against other benchmarks that we could potentially utilize.

2.2. Systemic risk analysis

For the purpose of facilitating xAI's systemic risk acceptance determination, xAI analyses each identified risks based on the following:

(1) **model-independent information**: xAI collects information relevant to the specific risk assessment, using methods such as web searches, literature reviews, market analysis, reviews of training data, historical incident data and incident databases, forecasting of general trends, expert interviews and/or panels and/or lay interviews, surveys, community consultations, or other participatory research methods investigating, e.g. the effects of models on natural persons, including vulnerable groups.

(2) **conducting model evaluations**: xAI conducts state-of-the-art model evaluations relevant to the systemic risk to assess the model's capabilities, propensities, affordances, and/or effects, which may include Q&A sets, task-based evaluations, benchmarks, red-teaming and other methods of adversarial testing, human uplift studies, model organisms, simulations, and/or proxy evaluations for classified materials.

(3) **modelling the systemic risk**: xAI conducts systemic risk modelling for relevant systemic risks by using state-of-the-art risk modelling methods, building on the systemic risk scenarios in the risk identification process and taking into account the information gathered.

(4) **estimating the systemic risk**: xAI estimates the the probability and severity of harm for the systemic risk using state-of-the-art risk estimation methods and taking into account the information gathered throughout the risk management process.

(5) **conducting post-market monitoring**: xAI conducts appropriate post-market monitoring to gather information relevant to assess whether the risk could be determined to not be acceptable and to inform whether a model update is necessary. Post-market monitoring may include collecting end-user feedback; providing (anonymous) reporting channels; providing (serious) incident reporting forms; monitoring software repositories, known malware, public forums, and/or social media for patterns of use.

2.3. Systemic risk acceptance determination

xAI applies a systemic risk acceptance criteria to each identified risk, incorporating a margin of security, to determine whether each identified systemic risk and the overall systemic risk are acceptable and, therefore, to proceed with the development, the making available on the

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[5] and inflict widespread harms upon the public.

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

market, and/or the use of the model. These evaluations are performed precedent to the wider deployment of our models, and are a precondition for release of our models for public use.

xAI uses rigorous evaluations to determine whether our models present systemic risks that remain within acceptable levels and to assess residual risk. Our consideration of residual risk takes account of the scale and probability of harm, and is evaluated in light of the sufficiency of implemented safeguards proportionate to the level of risk. In assessing whether to accept residual risk, we review the risk tiers for each systemic risk category and incorporate appropriate safety margins.

When analyzing whether a threshold has been reached, we also take into account additional evidence to validate the findings of evaluations, such as human expert red-teaming. These results help determine whether the threshold has been reached. The final decision also reflects an overall judgment based on all the available evidence, including how robust and reliable the evaluation methods were.

xAI will only proceed with the development, the making available on the market, and/or the use of the model, if the systemic risks stemming from the model are determined to be acceptable. In all scenarios, xAI will take appropriate measures to ensure the systemic risks stemming from the model are and will remain acceptable prior to proceeding. In particular, xAI may:

(1) to mitigate risks, employ 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;
(2) include 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
(3) conduct another round of systemic risk identification, analysis and systemic risk acceptance determination.

2.3. Safety mitigations

xAI implements appropriate safety mitigations to address the systemic risks stemming from our models, taking into account the model's release and distribution strategy, focusing 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).

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 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 malicious actor's progress through them.

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[6] 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):**[7] 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:**[8] A tendency toward excessive flattery or other sycophantic behavior has been observed in some production AI systems,[9] possibly resulting from directly 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.

---

[6] Taken out of context: On measuring situational awareness in LLMs
[7] The MASK Benchmark: Disentangling Honesty From Accuracy in AI Systems
[8] Towards Understanding Sycophancy in Language Models
[9] Sycophancy in GPT-4o: what happened and what we're doing about it

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.

xAI's objective is for our models to comply with their training, 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, as appropriate:

- **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.
- **Filters:** Applying classifiers to verify safety when a model is queried regarding topics of CRBN risks

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.

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.

2.4. Security mitigations

xAI will maintain a documented Security Goal that identifies the threat actors against which mitigations are designed to protect against, including sophisticated non-state actors, insider threats, state-sponsored actors and other foreseeable actors. The Security Goal will be reviewed at least every year.

xAI has implemented appropriate information security standards, adopted based on the NIST 800-171 Rev.3 framework and supported by SOC 2 Type II evaluations. Mitigations include

3. Safeguard Implementation

xAI trains its models to be honest and have values conducive to controllability, such as recognizing and obeying an instruction hierarchy.[10] 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 FAIF.

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

1. **Frontier AI Framework adherence:** Regularly review our adherence with this FAIF. Internally, we 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.

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

encryption of model weights, role-based access controls, and real-time monitoring to prevent unauthorized transfer. 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.[3]

Security for our AI services and processing systems is structured around comprehensive technical and organizational measures designed to protect the security, confidentiality, integrity, and availability of personal data and user content, along with encryption, access restrictions, and personnel safeguards. Access is tightly managed through least-privilege principles, role-based authorization, and regular reviews.

3. Incident reporting

xAI has measures in place to detect, investigate and remediate, as appropriate, serious AI safety incidents.

To foster accountability, we integrate the approach of designating risk owners, including assigning responsibility for proactively mitigating identified risks and monitoring for critical incidents or imminent threats, which may be identified through various channels, including:

- Red-teaming and internal testing;
- Real-time monitoring, telemetry, and alerting of threshold breaches via internal tooling;
- Monitoring and alerting of public comments from the X platform;
- Employee escalation;
- Feedback from end-users, downstream modifiers, downstream providers and other third parties through xAI's appropriate reporting channels.

When critical incidents or imminent threats are identified, xAI's will investigate the causes and effects of the incidents and, if necessary, take action to mitigate and contain incidents, and implement long-term solutions:

1. 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.

2. If we determine that allowing a system to continue running would materially and unjustifiably increase the likelihood of a risk, we may temporarily fully shut down the relevant system until we have developed a more targeted response.

3. We may perform a post-mortem of the event after it has been resolved, focusing on any areas where changes to systemic factors 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.

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

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. Practices include encryption of model weights, role-based access controls, and real-time monitoring to prevent unauthorized transfer. 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.[11]

4. Governance Approach

To foster accountability, we integrate the approach of designating risk owners, including assigning responsibility for proactively mitigating identified risks. Risk owners are also responsible for periodic audits to enforce framework implementation. Risk owners are also responsible for monitoring for critical incidents or imminent threats, which may be identified through:

- Red-teaming and internal testing;
- Real-time monitoring, telemetry, and alerting of threshold breaches via internal tooling;
- Monitoring and alerting of public comments from the X platform.

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.

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.

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

4. When reportable under applicable laws and regulations, xAI will provide the relevant authorities with a copy of the incident report within the required deadlines

4. Governance

xAI maintains internal governance structures and practices designed to meet the requirements of applicable laws and ensure the implementation of the processes in this FAIF. xAI's internal governance practices include managing risks across the lifecycle of our models and ongoing legal and compliance reviews to ensure that risk management functions adhere to this FAIF.

5. 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. Pre-deployment reviews include assessing benchmark results (e.g., WMDP scores) and mitigation effectiveness. For internal use, we review catastrophic risks like oversight evasion before extensive rollout. However, to ensure responsible deployment, this FAIF will be continually adapted and updated as circumstances change, before major new capabilities are launched, and in response to incidents. 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.

xAI Data Disclosure

This data disclosure is issued pursuant to California's AB-2013.

Grok is pretrained with a data recipe that includes publicly available Internet data, data produced by third parties for xAI, data from users (with the exception of Grok 1) or contractors, and internally generated data.

xAI aims to build AI models that are maximally truth-seeking, understand the true nature of the universe, and accelerate human scientific discovery, and xAI's use of its datasets is intended to further those purposes.

xAI's AI models were trained on datasets and dynamic datasets containing trillions to tens of trillions of tokens.

Grok is pretrained with a data recipe that includes publicly available Internet data, data produced by third parties for xAI, data from users (with the exception of Grok 1) or contractors, and internally generated data. In addition to pre-training, our recipe uses a variety of reinforcement learning techniques—human feedback, verifiable rewards, and model grading—along with supervised finetuning of specific capabilities.

Our datasets include data in the public domain and datasets that xAI has the necessary rights to use for training purposes.

xAI has the necessary rights to use the datasets it uses for training purposes, including because certain datasets were purchased or licensed.

Training datasets may incidentally include personal information, and their use is subject to compliance with applicable laws and regulations.

Training datasets may incidentally include aggregate consumer information, and their use is subject to compliance with applicable laws and regulations.

xAI cleans, processes or modifies datasets to facilitate pre-training, conduct reinforcement learning, and supervise finetuning.

Datasets were collected at various times since xAI was founded in March 2023. Data collection is ongoing.

Grok 1 began training on or about August 2023; Grok 1.5 began training on or about August 2023; Grok 2 began training on or about February 2024; Grok 3 began training on or about September 2024; Grok 4 began training on or about September 2024; Grok Code Fast 1 began training on or about September 2024; Grok 4 Fast began training on or about September 2024; and Grok 4.1 began training on or about May 2025.

xAI uses synthetic data generation in the development of its AI models in order to improve its AI models, including in reinforcement learning, finetuning, and post-training.