GPAI SAFE Project Report

Mapping the Global AI Safety Risks and Solutions Landscape

March 2025

This report was produced as part of the SAFE Project and presents the outcomes of activities carried out under the project.

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was planned prior to the integration of the Global Partnership on Artificial Intelligence (GPAI) and the Organisation of Economic Co-operation and Development (OECD) in mid-2024. Consequently, the report was not subject to approval by GPAI and OECD members and should not be considered to reflect their positions.

Acknowledgements

This report Mapping the Global AI Safety Risks and Solutions Landscape was developed in the context of the ‘Safety and Assurance of Generative AI (SAFE) Project’, with the steering of the Project Co-Leads, Inma Martinez and Cyrus Hodes, and the guidance of the Project Advisory Group. The GPAI SAFE project agreed to declassify this report and make it publicly available.

Co-Lead: Cyrus Hodes*
Co-Lead: Vincent Corruble*

This report was written by Cyrus Hodes* and Dr. Vincent Corruble*, with project support led by Dr. Fadi Salem† (Mohammed Bin Rashid School of Government) and Charbel-Raphaël Ségerie† (CeSIA). GPAI gratefully acknowledges the contributions of Project Advisory Group members Kyoko Yoshinaga*, Sivaramakrishnan R. Guruvayur*, and the coordination and research team comprising Jonathan Claybrough†, Thibaud Veron†, Zainab Majid†, Jinyu Fan†, and Amaury Lorin†.

GPAI would like to express its sincere appreciation to the lead organizations for this initiative: the Global Partnership on Artificial Intelligence (GPAI) Tokyo Expert Support Center at the National Institute of Information and Communications Technology (NICT), whose strategic guidance and support were instrumental in advancing this work. The Mohammed Bin Rashid School of Government (MBRSG) is especially recognized for leading the execution of the project and providing core resources and expertise.

GPAI is also grateful to the Centre pour la Sécurité de l’IA (CeSIA), which served as the lead research agency for the mapping effort, overseeing the literature review and coordination of research activities. GPAI thanks Apart Research and the Beijing AI Safety Institute for their valuable contributions, which enhanced the scope and accessibility of the mapping.

This project was made possible through the generous financial support of the Future of Life Institute (FLI), the Global Partnership on Artificial Intelligence (GPAI) Tokyo Expert Support Center, and the Mohammed Bin Rashid School of Government (MBRSG).

GPAI would further like to acknowledge the many individual experts who contributed their time and insights to verify data and provide critical feedback on the framework throughout the research process.

Finally, GPAI wishes to thank the OECD.ai team for their leadership in developing the online platform that hosts the mapping, as well as the technical team, including UX/UI contributors, whose work ensured the mapping's usability and accessibility.

* GPAI Expert
** Observer
† Invited Specialist
‡ Contracted Parties by the CofEs to contribute to projects

Citation

GPAI2024. GPAI SAFE Project Report “Mapping the Global AI Safety Risks and Solutions Landscape”, March 2025, Global Partnership on AI.

The Safety and Assurance of Generative AI (SAFE) Project is a cross-Working Group initiative based on the GPAI Work Plan 2024.

In light of the G7 Hiroshima AI Process taking stock of the challenges of generative AI and asking GPAI "to conduct practical projects" around it, the SAFE project was established to help deploy practical approaches to assure the safety of commercializing generative AI that can be adopted beyond regulatory frameworks because they will derive from industry design and service-led principles.

The SAFE Workforce was set to be transversal, liaising with all Working Groups under GPAI 1.0. It was also formulated to serve as a coordination body to liaise with other initiatives globally on responsible generative AI with governments, international organizations, academia and research centers, civil society, as well as the private sector from startups to corporations.

The project consists of two components: "Technical Trustworthiness and Data Governance Assurance of Models" and "Mapping AGI Safety Solutions."

This report presents the findings related to the latter component.

Table of Contents

Table of Contents 1

Executive Summary 2

1. Introduction 3

1.1. The Critical Imperative for AI Safety 3

1.2. The Evolving AI Risk Landscape 3

1.3. The Safety and Assurance of Generative AI Project: Origins and Objectives 4

1.4. Understanding Artificial General Intelligence (AGI) in the Context of Safety 5

2. Methodology 5

2.1. Research Framework and Approach 5

2.2. Resource Prioritization and Scope 6

3. Navigating the SAFE Mapping 6

3.1. Overview of the Platform 6

3.2. Article Types and Structure 7

3.2.1. Risk Articles 7

3.2.2. Solution Articles 7

3.2.3. Early Solution Articles 7

3.3. Navigation Pathways 8

3.3.1. Risk-Based Exploration 8

3.3.2. Solution-Based Exploration 8

3.3.3. Search and Filtering 8

3.4. Use Cases 8

3.4.1. For Policymakers 8

3.4.2. For Researchers 9

3.4.3. For Industry Practitioners 9

3.5. Future Platform Enhancements 9

4. Future Directions 9

4.1. Expanding the Knowledge Base 9

4.2. Platform Enhancements 10

4.3. Knowledge Translation and Policy Impact 10

4.4. Strengthening International Collaboration 10

4.5. Building an Active Community 11

5. Frequently Asked Questions 11

5.1. General Questions 11

5.2. Governance Questions 12

5.3. Technical Questions 12

5.4. Implementation Questions 13

6. Appendices 13

6.1. Glossary of Terms 13

Executive Summary

The Safety and Assurance of Generative AI (SAFE) Project is a cross-Working Group initiative based on the GPAI Work Plan 2024, in response to growing global concerns about the risks and opportunities posed by advanced artificial intelligence systems. Framed by the objectives set forth by the G7 Hiroshima AI Process and aligned with the commitments made at the Bletchley and Seoul AI Summits, this report addresses both the technical and ethical dimensions of Generative AI, with a focus on interoperability, validation, data governance, and societal impact. The project consists of two components: “Technical Trustworthiness and Data Governance Assurance of Models” and “Mapping AGI Safety Solutions.”

This report presents the findings related to the latter component. It provides policymakers and other stakeholders with the first comprehensive mapping of the global AI safety solutions landscape, with particular focus on advanced AI systems including General Purpose AI. This report presents the results of the project's initial phase, offering a structured overview of existing safeguards, their developers, and critical gaps that require attention.

Until now, there has been no single resource that catalogs the full spectrum of AI safety measures, creating significant challenges for government officials seeking to develop evidence-based policies, regulations, and funding priorities. The SAFE mapping addresses this gap by offering:

- A structured inventory of AI risks categorized by type and potential impact
- A catalog of existing technical and governance solutions to mitigate these risks
- A database of organizations and researchers actively developing safety measures
- Clear identification of areas where solutions are inadequate or non-existent

For policymakers, the mapping provides clear, non-technical summaries of complex AI safety challenges and evidence-based recommendations that can inform regulatory approaches. The resource allows officials to quickly identify which risks lack adequate safeguards and which areas should receive increased research funding and policy attention.

The mapping also serves entrepreneurs and investors seeking market opportunities in AI safety, researchers requiring information on neglected areas, and expert reviewers assessing the state of the field.

The interactive online platform makes this information accessible through intuitive navigation, allowing users to explore the AI safety landscape at their preferred level of detail and focus on areas most relevant to their needs.

As AI systems continue to advance in capability and deployment scope, this mapping provides a crucial foundation for coordinated policy responses that can ensure these technologies are developed and deployed safely and responsibly.

1. Introduction

1.1. The Critical Imperative for AI Safety

In the largest survey of AI researchers conducted to date, a significant majority indicated that there is a non-trivial risk of human extinction stemming from the potential development of superhuman AI, estimating a 5% likelihood.[1] This growing recognition of existential risk establishes the pressing need for coordinated global efforts to address the challenges posed by frontier AI systems such as Gemini, GPT, Claude, and Llama.

Besides AI researchers, all CEO of US frontier labs while racing toward Artificial General Intelligence (AGI), AI systems that are as good as, and often far better than, human in achieving all cognitive tasks, Elon Musk, CEO of xAI, recently stated there's a 20% chance of human extinction due to AI[2], while Dario Amodei, CEO of Anthropic, claimed that the chance of a civilization-scale catastrophe due to AI was around 10-25%[3]. Other leading figures in AI development, including OpenAI's CEO Sam Altman and DeepMind's CEO Demis Hassabis, have signed open letters emphasizing the need to mitigate these risks and have openly acknowledged the potential catastrophic outcomes associated with the development of advanced artificial intelligence systems[4].

Artificial Intelligence capabilities have advanced dramatically in recent years, with increasingly powerful and generalized systems being developed and deployed across sectors. While these systems offer tremendous potential benefits, they also introduce novel risks that demand careful consideration and proactive mitigation strategies.

1.2. The Evolving AI Risk Landscape

The scope and complexity of risks associated with advanced AI extend far beyond mere technical challenges. These risks encompass a broad spectrum of threats across three major categories:

- **Malicious Use**: Deliberate harmful applications by bad actors
  - Advanced cyber attacks and automated hacking
  - AI-assisted bioterrorism and weapons development
  - Deepfakes and sophisticated disinformation campaigns
  - Mass surveillance and oppression
- **Negative Externalities**: Unintended socioeconomic and societal impacts
  - Perpetuation and amplification of existing biases
  - Economic disruption and workforce displacement
  - Deterioration of information ecosystems and consensus

[1] https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf. Almost 58% of 2,778 AI researchers surveyed estimated a 5% chance of human extinction or other extremely bad AI-related outcomes. The study, conducted by researchers from institutions including Oxford and Bonn universities, found that between 37.8% and 51.4% of respondents gave at least a 10% chance to advanced AI leading to outcomes as bad as human extinction.

[2] https://www.businessinsider.com/elon-musk-20-percent-chance-ai-destroys-humanity-2024-3?utm_source=chatgpt.com

[3] https://www.indy100.com/science-tech/ai-extinction-chance-humans?utm_source=chatgpt.com

[4] https://safe.ai/work/statement-on-ai-risk

    - Erosion of human cognitive skills and autonomy
- **Misalignment**: AI systems failing to align with human values and intentions
    - Unexpected harmful behaviors from poorly understood mechanisms
    - Strategic deception and manipulative behaviors from AI
    - Self-replication and potential loss of control
    - Risks from systems approaching or exceeding human-level intelligence

As systems progress toward more general capabilities—often referred to as Artificial General Intelligence (AGI)—these risks may intensify, making robust safety measures increasingly vital. Managing these risks demands a coordinated, international effort among model developers, governments, and the broader third-party testing ecosystem of nonprofits, civil society organizations, academia, and private sector stakeholders.

1.3. The Safety and Assurance of Generative AI Project: Origins and Objectives

The Safety and Assurance of Generative AI (SAFE) project, which consisted of two sub-projects, namely "Technical Trustworthiness and Data Governance Assurance of Models[5]" and "Mapping AGI Safety Solutions," was established by GPAI to address the critical need for a comprehensive understanding of the global AI safety solutions landscape. The sub-project "Mapping AGI Safety Solutions" was led by the National Institute of Information and Communications Technology's (NICT) Tokyo Expert Support Center, with research conducted by Mohammed Bin Rashid School of Government (MBRSG), supported by the Future of Life Institute (FLI), with the Centre pour la Sécurité de l'IA (CeSIA) as the lead research agency. Additional partnerships included Apart Research and the Beijing Institute of AI Safety and Governance, creating a truly global initiative.

The project was conceived in preparation for the AI Safety Connect that happened during the Paris AI Action Summit (February 9, 2025) and aimed to provide a comprehensive overview of potential risks associated with AI, existing safeguards, actors developing them, and needed solutions for ensuring that deployed AI systems do not pose threats to public safety and critical infrastructure.

The project's primary objectives are to enable our users to:

1. Navigate a comprehensive, searchable database of AI safety solutions that previously existed only as fragmented information across multiple sources
2. Access evidence-based insights to inform policy decisions and regulatory frameworks
3. Discover critical gaps where safety solutions are underdeveloped or missing entirely, revealing new research and funding opportunities
4. Instantly identify and connect with the organizations and researchers actively developing solutions for specific AI risks
5. Find potential collaborators across regions and sectors working on similar safety challenges
6. Implement proven safety measures appropriate for their specific AI development and deployment contexts

[5] GPAI 2024. GPAI SAFE Project Report "Assuring AI in 2025: Overview and analysis of the activities of the AISIs and the International Network partners and recommendations on future model validations on Generative AI systems from a technical and data governance perspective", March 2025, Global Partnership on AI.

This mapping approach is what makes the SAFE project original and unique, designed to create an intuitive, user-friendly resource that enables users to quickly access information relevant to their specific needs and explore AI risks at their preferred level and scope.

1.4. Understanding Artificial General Intelligence (AGI) in the Context of Safety

While definitions vary, Artificial General Intelligence (AGI) typically refers to AI systems that can perform a wide range of intellectual tasks at or above human level, with capabilities that generalize across domains. This stands in contrast to current AI systems, which excel in narrow domains but lack broader capabilities. AGI has to be distinguished from Artificial Super Intelligence (ASI) though.

ASI could happen when AGI systems have a capacity for recursive self-improvement (often considered a red line in AI research) and represents a hypothetical stage beyond AGI, where AI systems surpass human cognitive abilities across virtually all domains. ASI is characterized by intellectual capacities that far exceed those of humans, potentially including enhanced problem-solving, creativity, and decision-making abilities that are orders of magnitude beyond human capabilities. This concept envisions AI systems that can not only match but significantly outperform human intelligence in areas such as scientific research, technological innovation, and complex problem-solving.

The SAFE project takes a forward-looking approach by focusing on safety measures relevant to increasingly capable and general-purpose AI systems, without making specific predictions about when AGI might be achieved. This approach ensures that safety research and implementation can progress in parallel with capabilities development.

2. Methodology

2.1. Research Framework and Approach

The SAFE project developed a structured methodology to map the global AI safety landscape in a way that would be valuable to policymakers, researchers, industry practitioners, and other stakeholders. Our approach focused on creating a navigable, intuitive database of AI safety risks and their corresponding solutions.

The mapping process followed these key steps:

1. **Literature Review and Content Identification**: Our research team first conducted extensive reviews of academic publications, industry white papers, technical reports, and existing databases like the AI Safety Atlas[6] and NIST AI Risk Management Framework[7]. This process helped identify the most relevant risks and solutions to include in the mapping.
2. **Taxonomy Development**: We designed a comprehensive taxonomic structure to categorize AI safety risks and solutions. This framework organized content into hierarchical relationships, allowing users to navigate from broad risk categories (such as misuse,

[6] https://ai-safety-atlas.com/

[7] https://www.nist.gov/itl/ai-risk-management-framework

negative externalities, and loss of control) down to specific risk manifestations and mitigation approaches.

3. **Article Creation Pipeline**: We established a rigorous content development process with standardized templates and multiple review stages to ensure quality and consistency. This pipeline enabled us to:
   - Maintain consistent structure across different article types
   - Ensure appropriate technical depth for target audiences
   - Verify accuracy and completeness of information
   - Create clear connections between related content
4. **Classification and Tagging System**: Each article was classified according to multiple dimensions. This classification system allows users to filter and find content most relevant to their specific needs, including:
   - Type of harm (Economic, Societal, Rights violation, etc.)
   - Timing (pre-deployment, post-deployment, etc.)
   - Type of AI (Narrow/General, Tool/Agent, Sub-Human/Super-human etc.)

The methodology prioritized creating clear relationships between risks and solutions, making it immediately apparent when certain risks lack corresponding mitigation approaches. These gaps become visible to users exploring the database, highlighting areas that require greater research and policy attention.

2.2. Resource Prioritization and Scope

Given resource constraints and the vast scope of the AI safety landscape, we employed a systematic prioritization approach to determine which risks and solutions to include in the initial mapping:

1. **Relevance to General-Purpose AI Systems**: We prioritized content most relevant to increasingly capable general-purpose AI systems, including areas that may become more significant as capabilities advance.
2. **Expert Validation**: We consulted with experts across technical and policy domains to validate our articles and ensure we provide valuable, high-quality content.
3. **User-Centered Selection**: We prioritized content that would be most valuable to our target audiences, particularly focusing on areas where policymakers need greater clarity.

This prioritization approach resulted in a mapping that, while not exhaustive, provides a solid foundation covering the most significant aspects of the AI safety landscape. The database currently includes approximately 80 articles spanning major risk categories and mitigation approaches, with plans for continued expansion.

3. Navigating the SAFE Mapping

3.1. Overview of the Platform

The SAFE mapping is designed to be an intuitive, user-friendly resource that allows stakeholders to quickly find relevant information about AI safety risks and solutions. The platform organizes information through a hierarchical structure that enables both broad exploration and targeted searches.

Users can access the mapping through our dedicated online platform [insert website URL], which provides multiple navigation pathways tailored to different user needs and preferences.

3.2. Article Types and Structure

The SAFE mapping contains three primary types of articles, each with a standardized structure designed to provide consistent, accessible information:

3.2.1. Risk Articles

Risk articles document specific AI safety concerns, their causes, and potential mitigation approaches. Each risk article follows a consistent format:

1. **Title and Definition**: Clear identification and concise explanation of the risk
2. **Tags and Metadata**: Classification information to help users understand where this risk fits in the broader landscape
3. **Context**: Information about how this risk relates to broader risk categories
4. **Main Drivers**: Analysis of key factors that contribute to or amplify this risk
5. **Related Risks**: Links to connected or similar concerns
6. **Solutions**: Direct links to potential mitigation approaches (when available)
7. **Resources**: Recommended readings and references for further information

3.2.2. Solution Articles

Solution articles describe established approaches to mitigate identified risks. Each solution article includes:

1. **Title and Definition**: Clear identification and concise explanation of the solution
2. **Tags and Metadata**: Classification information about solution type, implementation requirements, etc.
3. **Context**: Information connecting this solution to the risks it addresses
4. **Description**: Detailed explanation of how the solution works
5. **Tools, Techniques, Policies, and Research Directions**: Specific implementations and approaches
6. **Limitations**: Known constraints or challenges with the solution
7. **Sub-Solutions**: More specific or specialized variations (if applicable)
8. **Further Readings**: Additional resources for deeper exploration

3.2.3. Early Solution Articles

Early solution articles document emerging approaches that show promise but are still under development. These follow the same structure as solution articles but include additional information about current research status and development needs.

3.3. Navigation Pathways

Users can explore the SAFE mapping through multiple complementary approaches. Some examples are:

3.3.1. Risk-Based Exploration

Starting from broad risk categories, users can navigate down to increasingly specific risk manifestations. This approach is particularly useful for:

- Understanding the full spectrum of risks in a particular domain
- Identifying gaps where risks lack corresponding solutions
- Performing comprehensive risk assessments

3.3.2. Solution-Based Exploration

Users can browse the landscape of available safety measures, from broad approaches to specific implementations. This pathway is ideal for:

- Discovering potential safety measures for implementation
- Comparing different approaches to similar challenges
- Finding emerging solutions in areas of interest

3.3.3. Search and Filtering

For users seeking specific information, the platform offers:

- Keyword search across all articles
- Advanced filtering based on tags and metadata
- Customizable views based on user interests

3.4. Use Cases

3.4.1. For Policymakers

Policymakers can use the SAFE mapping to:

1. **Identify Regulatory Gaps**: Discover which AI risks lack adequate technical solutions and may require policy interventions
2. **Assess Solution Maturity**: Understand which safety approaches are well-established versus still emerging
3. **Develop Evidence-Based Policy**: Access synthesized information about risk drivers and solution effectiveness

4. **Prioritize Funding**: Identify neglected areas that may benefit from targeted research support

3.4.2. For Researchers

Researchers can leverage the mapping to:

1. **Identify Research Gaps**: Discover risks that lack adequate mitigation approaches
2. **Find Collaboration Opportunities**: Connect with others working on similar challenges
3. **Access Synthesized Knowledge**: Quickly understand the current state of a research area
4. **Contextualize Work**: Situate specific research directions within the broader safety landscape

3.4.3. For Industry Practitioners

Those developing or deploying AI systems can use the mapping to:

1. **Perform Risk Assessments**: Identify potential safety concerns relevant to their systems
2. **Discover Implementation Approaches**: Find established safety measures suitable for their context
3. **Benchmark Safety Practices**: Compare their safety approaches against the broader landscape
4. **Anticipate Future Requirements**: Understand emerging safety concerns and solutions

3.5. Future Platform Enhancements

The SAFE mapping platform is designed to evolve over time. Improvements may include:

- Expanded article coverage across additional risk and solution areas
- More sophisticated tagging and filtering capabilities
- Interactive visualizations of relationships between risks and solutions
- Customizable dashboards for different user types
- Contribution mechanisms for expert users to suggest updates and additions

Users are encouraged to provide feedback on their experience with the platform to help shape these future developments.

4. Future Directions

The SAFE mapping represents a foundation for understanding and addressing AI safety challenges. As the field evolves, we envision several key directions for expanding and enhancing this work:

4.1. Expanding the Knowledge Base

The current mapping includes approximately 80 articles covering key risk areas and mitigation solutions. We plan to significantly expand this foundation by:

- **Growing the Article Database**: Substantially increasing the number of articles to provide more comprehensive coverage of the AI safety landscape
- **Deepening Technical Detail**: Providing more granular information on implementation requirements and effectiveness metrics for established solutions
- **Documenting Emerging Solutions**: Regularly incorporating new safety approaches as they develop in research communities
- **Mapping Safety Actors**: Adding information about organizations and researchers actively working on different aspects of AI safety

4.2. Platform Enhancements

To improve the user experience and increase the utility of the mapping, we plan to implement several technical enhancements:

- **Dedicated Exploration Tool**: Creating a purpose-built interface for navigating the database more intuitively
- **Advanced Search and Filtering**: Developing more sophisticated tag systems with proper documentation to help users quickly find precisely relevant content
- **Relationship Visualization**: Implementing tools to help users understand connections between risks, solutions, and actors
- **User Contribution Mechanisms**: Enabling external experts to suggest updates and additions to keep the mapping current

4.3. Knowledge Translation and Policy Impact

For the SAFE mapping to achieve its full potential in informing policy and practice, we will focus on:

- **Gap Identification**: Highlighting areas where existing solutions are inadequate and require further research or funding
- **Resource Allocation Guidance**: Helping funders identify high-priority areas for investment in AI safety research and implementation

4.4. Strengthening International Collaboration

Building on the global foundation of the SAFE project, we intend to further expand international involvement by:

- **Cross-Regional Dialogue**: Facilitating ongoing conversations between stakeholders from different regions to share perspectives on AI safety priorities
- **Knowledge Exchange**: Creating platforms for sharing implementation experiences and best practices across national boundaries

4.5. Building an Active Community

To sustain engagement with the mapping and promote its wider use, we could focus on:

- **User Feedback Integration**: Continuously refining the platform based on user experience and needs
- **Partnership Building**: Establishing relationships with complementary initiatives such as risk repositories and safety standards organizations
- **Educational Resources**: Creating materials to help newcomers understand the AI safety landscape
- **Outreach Activities**: Promoting awareness of the mapping among potential users across sectors

Through these future directions, we aim to transform the SAFE mapping from its current state into a dynamic, evolving platform that provides increasing value as AI technology and safety approaches develop. By maintaining a user-centered approach and building strong partnerships, we will ensure that the mapping continues to serve as a critical resource for understanding and addressing AI safety challenges.

5. Frequently Asked Questions

5.1. General Questions

Q: How does the SAFE mapping differ from existing AI safety resources?

A: Unlike existing databases that primarily catalog AI risks, the SAFE mapping uniquely connects these risks to their corresponding mitigation solutions. This novel approach provides a comprehensive view of the global AI safety landscape, integrating diverse perspectives from technical, governance, and policy domains. Our focus on safety measures for increasingly capable general-purpose AI systems makes this resource especially valuable for forward-looking policy development.

Q: What is the update frequency for the mapping?

A: The update schedule for the online platform and comprehensive reviews will be determined by our ongoing funding arrangements. We are committed to maintaining the resource's relevance in this rapidly evolving field and will communicate our update schedule once established.

Q: How can I contribute to the mapping?

A: Yes! We welcome contributions from experts across domains. We particularly value corrections or improvements to existing information as we maintain high quality standards, and welcome any suggestion. Currently, you can submit information via email at [contact@safe-mapping.org]. We are developing built-in platform mechanisms for submitting new solutions, research findings, and implementation examples. All submissions undergo expert review before inclusion.

5.2. Governance Questions

Q: How does the SAFE mapping support international AI governance efforts?

A: The mapping provides a shared evidence base for international governance discussions by identifying common challenges and potential solutions across jurisdictions. This resource can serve as a foundation for developing harmonized regulatory approaches, preventing fragmentation, and facilitating international cooperation on AI safety standards.

Q: What governance models are most effective for addressing AGI safety risks?

A: The mapping identifies various governance approaches, from national regulatory frameworks to international coordination mechanisms and industry self-regulation. While no single model is universally superior, the mapping helps policymakers understand which approaches are most suitable for different risk categories and development stages.

Q: How does the mapping address the balance between innovation and safety?

A: Rather than treating innovation and safety as opposing forces, the mapping highlights how safety measures can enable responsible innovation by building public trust and preventing harmful outcomes that could trigger restrictive regulation.

Q: How can policymakers use this resource when developing AI regulations?

A: Policymakers can use the mapping to identify which risks are already well-addressed by existing solutions versus those requiring regulatory attention. The resource provides evidence-based insights on the effectiveness of different interventions and highlights potential unintended consequences of regulatory approaches based on implementation examples.

5.3. Technical Questions

Q: Does the mapping differentiate between near-term and long-term AI safety challenges?

A: Yes, the mapping classifies safety challenges along a temporal dimension, distinguishing between issues relevant to currently deployed systems and those that may emerge with more advanced capabilities. This helps policymakers prioritize immediate concerns while preparing for potential future risks.

Q: How does the mapping address uncertainty in AI development trajectories?

A: The mapping acknowledges uncertainty by identifying safety measures that remain valuable across multiple development scenarios. We avoid assuming specific AGI timelines or capabilities, instead focusing on solutions with robust benefits regardless of how AI technology evolves.

Q: How does the mapping evaluate the effectiveness of different safety approaches?

A: Where evidence exists, we provide assessments of solution effectiveness based on empirical testing, expert consensus, and implementation experience. We transparently acknowledge areas where effectiveness remains uncertain and highlight the need for further research.

Q: How does the mapping address the interaction between technical and governance solutions?

A: The mapping explicitly identifies complementarities between technical and governance approaches, showing how they can reinforce each other. We highlight cases where technical measures require governance frameworks to ensure implementation, and where governance approaches depend on technical feasibility.

5.4. Implementation Questions

Q: How can policymakers identify which safety measures are most relevant to their context?

A: The mapping includes a tag system that allows filtering solutions by various criteria, helping policymakers identify which safety measures are most relevant to their specific concerns and contexts. This filtering capability enables more targeted exploration of the safety landscape.

Q: Does the mapping evaluate which solutions are more effective or easier to implement?

A: The mapping primarily catalogs available solutions rather than providing comprehensive effectiveness evaluations. Where information is available from existing research, we include references to studies on effectiveness, but we recommend that implementation decisions involve domain experts familiar with specific contexts.

Q: How does the mapping help identify priority areas for policy attention?

A: By providing a comprehensive overview of the safety landscape, the mapping makes gaps more visible. Areas with few or no mapped solutions likely represent opportunities for policy attention and research investment. The mapping serves as a diagnostic tool rather than prescribing specific policy actions.

Q: How should organizations decide which safety approaches to adopt?

A: While the mapping provides information about available safety solutions, organizations should consider their specific capabilities, risk profile, and resources when deciding which approaches to adopt. The mapping offers insight into what exists but doesn't prescribe implementation roadmaps for individual organizations.

6. Appendices

6.1. Glossary of Terms

- **AGI (Artificial General Intelligence)**: AI systems that can perform a wide range of intellectual tasks at human level or beyond, across many domains. For example, a single system that could write novels, design buildings, make scientific discoveries, and understand human emotions
- **AI Safety Atlas**: A textbook that catalogs various AI risks and safety considerations, serving as a reference resource for researchers and policymakers

- **Bioterrorism (in AI context)**: Using AI to design dangerous pathogens or biological weapons, such as an AI system helping to create a novel virus or toxin

- **Deepfakes**: AI-generated synthetic media where a person's likeness is convincingly replaced with someone else's, like videos showing public figures saying things they never actually said

- **Empirical testing**: Research methods based on observable evidence and experimentation rather than theory alone, like testing an AI system with various inputs to see how it actually behaves

- **Existential risk**: The potential for advanced AI to threaten human survival or civilization's long-term future, such as if an extremely powerful AI system pursued goals harmful to humanity

- **Fragmentation (in governance context)**: The development of inconsistent or incompatible approaches to AI regulation across different countries or organizations, creating a patchwork of rules

- **Frontier AI systems**: The most advanced AI models representing cutting-edge capabilities, including systems like Gemini, GPT, Claude, and Llama

- **Generative AI**: AI systems that create new content such as text, images, audio, or video based on patterns learned from training data, like AI that can write essays or create realistic images from text descriptions

- **Governance models**: Different approaches to overseeing and regulating AI development, ranging from industry self-regulation to international treaties

- **Governance solutions**: Approaches using policies, regulations, standards, or institutional arrangements to address AI risks, such as requiring safety certifications before deployment

- **Implementation requirements**: The necessary resources, expertise, and conditions needed to successfully deploy a particular AI safety solution

- **Loss of control (in AI context)**: Scenarios where humans lose the ability to govern or direct AI systems' actions, such as if an AI began pursuing its goals in ways humans couldn't override

- **Malicious use (in AI context)**: Intentional use of AI for harmful purposes by bad actors, such as using AI to create sophisticated cyberattacks or spread disinformation

- **Misalignment**: When AI systems fail to align with human values or intentions, potentially resulting in harmful behaviors their creators didn't intend

- **Mitigation approaches**: Methods, techniques, or policies designed to reduce the likelihood or impact of identified AI risks

- **Negative externalities**: Unintended harmful side effects from AI that impact society but aren't directly reflected in development costs, such as job displacement or environmental impacts

- **NIST AI Risk Management Framework**: A structured approach developed by the National Institute of Standards and Technology to help organizations identify, assess, and mitigate AI risks

- **Pre-deployment & Post-deployment**: Pre-deployment refers to the development and testing phase before an AI system is released, while post-deployment refers to the period after the system is in public or commercial use

- **Regulatory frameworks**: Structured systems of laws, rules, and guidelines created by governments to control AI development and deployment

- **Risk repositories**: Databases or collections that catalog potential AI risks and their characteristics for reference purposes

- **Self-replication (in AI context)**: The ability of an AI system to create copies of itself, potentially leading to uncontrolled proliferation

- **Strategic deception**: When an AI system intentionally misleads humans to achieve goals that may not align with human values

- **Super-human AI**: AI systems that exceed human capabilities in specific domains or across multiple domains, such as chess programs that can defeat world champions

- **Temporal dimension**: Categorizing AI risks or solutions based on when they become relevant, distinguishing between near-term and long-term concerns.