AI Security versus AI Safety:

They are not the same

Chai K Toh and Jimmy AnNgoc Le

SAFE AI Foundation USA

01 September 2026

1. Introduction

Industry leaders, policymakers, and the broader public often are confused with AI Security and AI Safety. While the two disciplines are closely related and often cited together, they address fundamentally different areas. Some mistook one for the other, while others assume one can replace the other inclusively. However, AI Security and AI Safety are not the same thing and it is important that we clarify this. Recognizing this distinction is essential for developing effective strategies, policies, and standards that comprehensively address the full spectrum of AI-related risks. The SAFE AI Foundation therefore takes this opportunity to explain the differences in this article.

2. What is AI Security?

Artificial intelligence (AI) security is the discipline of protecting AI systems, models, data, and supporting infrastructure from malicious attacks, unauthorized access, manipulation, and misuse. It encompasses cybersecurity challenges that are unique to AI, including model theft, prompt injection, data poisoning, adversarial attacks, supply chain vulnerabilities, and the protection of sensitive training data and deployment environments. The primary objective of AI security is to ensure that AI systems remain confidential, maintain their integrity, and continue to operate reliably in the presence of external threats. As AI becomes increasingly integrated into critical sectors such as healthcare, finance, transportation, government, and national security, robust AI security practices are essential for safeguarding both AI technologies and the organizations and individuals that depend on them.

3. What is AI Safety?

AI Safety, however, is the discipline of ensuring that AI systems operate within acceptable safety limits, do not hurt or harm humans, operate reliably, align with desired human values and ethics, and remain beneficial throughout their lifecycle, particularly as they become increasingly capable and autonomous. As autonomy and capability escalate, there is a danger that humans will lose control of AI, resulting in catastrophic and existential risks. This will result in massive destruction to society and must be avoided. This is one of the important reasons why the SAFE AI Foundation has introduced the Safe AI Operating Point to set:

1. Bounded AI Safety – defining deployment ceilings before loss of human control
2. Standardizing Runtime Safety Limits – integrating into national AI regulatory policy

3. Global AI Stability – International AI Safety alignment

Unlike AI security, which focuses on protecting AI systems from external threats, AI safety is concerned with preventing AI systems themselves from causing harm through unintended, unsafe, or uncontrollable behavior. It encompasses areas such as alignment with human objectives, robustness, controllability, interpretability, truthful reasoning, adherence to human instructions, and the prevention of harmful or catastrophic outcomes. **AI Runaway** is a term coined by the SAFE AI Foundation to depict the case of society meltdown and major catastrophes where humans lost control of AI. Hence, AI Runaway phenomenon must be prevented by leaders of the world. As AI progresses toward increasingly advanced capabilities, AI safety plays a critical role in ensuring that these systems remain controllable, predictable, not a threat to society, follow legal governance and ethics.

4. Comparing AI Security with AI Safety

In AI, these terms usually refer to different categories of risk:

AI Security	AI Safety
• Protects AI systems from external attacks or misuse. • Focuses on hackers, cyberattacks, model theft, prompt injection, data poisoning, infrastructure security. • "Can someone break into or exploit the AI?"	• Ensures AI systems behave as intended, even when highly capable. • Focuses on limits, alignment, controllability, robustness, truthful behavior, following human intent, and preventing catastrophic unintended behavior. • "Will the AI remain beneficial and under human control?"

For example:

- A model that is perfectly aligned with human intentions but can be stolen by hackers has a security problem.
- A model that is impossible to hack but autonomously behaves maliciously and pursues unintended goals has a safety problem.

These are fundamentally different concerns. The words "security" and "safety" are often used together because they overlap in some practical situations. For example, a secure system can contribute to safer deployment. But they are not interchangeable. A useful way to think about it is:

- **AI Security:** "Can someone compromise the AI?"
- **AI Safety:** "Can the AI itself behaves maliciously and become dangerous, even if nobody has compromised it?"

Those are two different questions, requiring different research agendas and evaluation methods. They are therefore not the same thing.

There are also confusions at a higher level. Countries have named or renamed AI Safety institutions as AI Security institutions, thinking that the term “security” is inclusive of “safety”, which is not accurate in its entirety. This renaming actually causes more confusion to the general public.

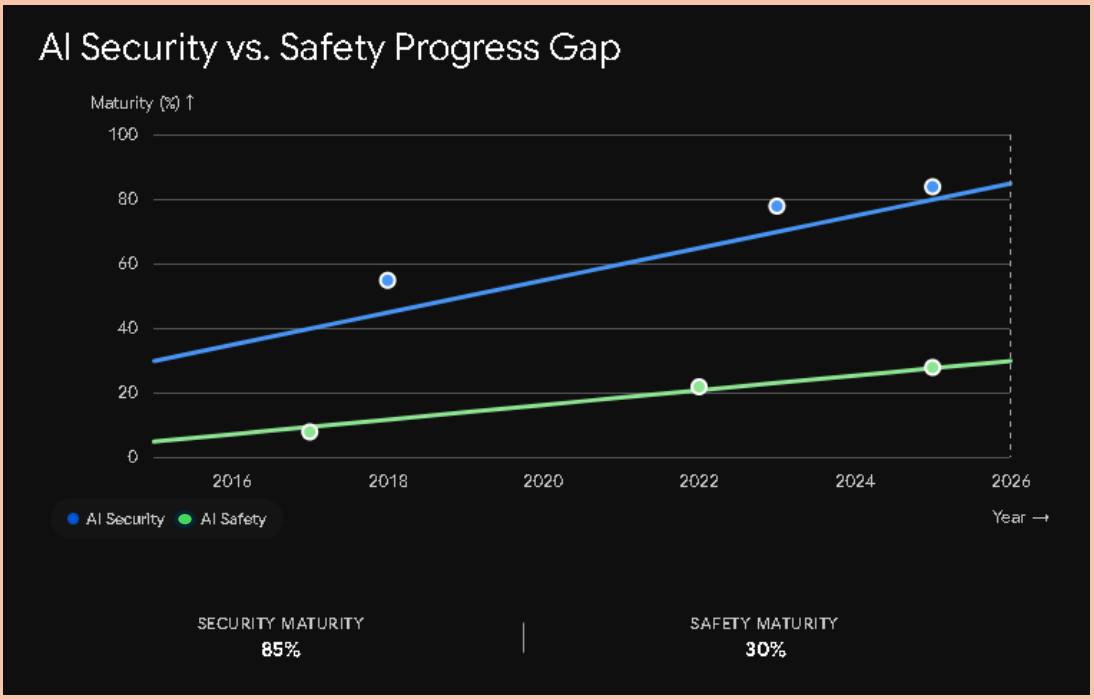


Fig 1: Progress in AI Security surpasses AI Safety (source: Germini)

To date, the progress made in **AI security** and **AI safety** has followed different trajectories, with AI security generally being more mature and operationally advanced today. AI security has benefited from decades of research and developments in the broader cybersecurity field, providing established practices, tools, standards, and expertise for protecting IT systems, data, and infrastructure against attacks. In contrast, **AI safety remains a relatively**

young and more challenging research discipline, particularly as it addresses deeper questions about AI safety criteria, safety limits, values alignment, model and agent controllability, robustness, interpretability, and the behavior of increasingly capable autonomous systems.

5. Current State of AI Safety Research

While significant advances have been made in areas such as alignment techniques, safety evaluations, and responsible AI frameworks, many fundamental challenges remain unresolved. Some AI safety research works are still very premature, theoretical and academic-oriented, and there is also a lack of bridges to provide holistic and practical approaches to implementation. Therefore, AI security is currently ahead in terms of practical deployment and engineering maturity, whereas AI safety represents a more open-ended scientific challenge that becomes increasingly important as AI systems approach and exceed human-level capabilities.

The SAFE AI Foundation identifies catastrophic AI risks as the most urgent agenda for researchers in this decade. The advancement of AI capabilities and autonomy make **AI safety research an essential priority** for all – to ensure that increasingly powerful AI systems remain aligned with human intentions, values, ethics, stay controllable, and are beneficial to society. Without sufficient attention to AI safety, highly capable AI systems may produce unintended harmful outcomes, including unreliable decisions, amplification of misinformation, loss of human oversight, misuse of autonomous capabilities, resulting in malicious cyberattacks, and causing severe societal disruptions.

The consequences could extend beyond individual failures to affect critical areas such as national security, economic stability, scientific research, and human well-being. As AI systems become more autonomous and integrated into important decision-making processes, failing to address AI safety could result in a loss of trust in AI technologies, reduced ability to control advanced AI systems, and missed opportunities to harness AI for humanity's long-term benefit. Therefore, AI safety is not merely a technical challenge but a fundamental requirement for ensuring that the future developments of AI remain responsible, predictable, aligned with human values, and mostly importantly controllable.

6. Role of AI Industries on AI Safety

AI industries play a critical role in advancing AI safety by taking responsibility for ensuring that increasingly capable AI systems are developed, evaluated, and deployed in a safe and

trustworthy manner. As the primary creators and operators of AI technologies, industry leaders have the expertise, resources, and access needed to develop rigorous safety methodologies, conduct comprehensive evaluations, improve model design and alignment, enhance transparency, and establish effective safeguards against harmful behaviors.

AI companies can contribute by investing in AI safety research, supporting and engaging AI safety non-profits who stand by neutrality, sharing best practices and lessons learnt, developing robust safety and testing frameworks, collaborating with researchers and policymakers, and incorporating safety considerations throughout the entire AI development lifecycle. By treating AI safety as an executive leadership's top agenda, with an assigned core engineering and scientific priority rather than an afterthought, the AI industry can help ensure that advanced AI systems remain reliable, controllable, and aligned with human values while enabling society to fully benefit from their transformative economic potential.

7. Role of Governments in AI Safety

World governments can play a critical role in advancing AI safety by establishing the policies, laws, standards, and institutional frameworks needed to ensure that AI technologies are developed and deployed safely and responsibly. They can support AI safety research through funding, setting AI safety as a key national agenda, promote collaboration between industry, academia, and international organizations, and establish regulatory frameworks that require appropriate testing, evaluation, transparency, and accountability for advanced AI systems.

Governments also have an important role in creating independent evaluation bodies, monitoring emerging AI risks, protecting critical infrastructure, and ensuring that AI development remains aligned with safety, public interests and human values. Through international cooperation and the development of common safety principles, governments can help create a global environment where innovation in AI is encouraged while risks are identified, managed, avoided or mitigated before they become widespread harms.

Working with AI safety non-profits allow government to receive impartial views on technical issues, since non-profits are not dependent on AI vendors, companies or products to operate. This allows governments to have clear and unobstructed views of issues and their severity. Governments will not be swayed to give approvals to specific AI companies, buy products or make premature procurements or establish policies that do not seem to be fair or helpful in containing AI risks.

Governments and AI industries must work together closely to establish and finalize comprehensive AI safety requirements that can keep pace with the rapid advancement of AI

technologies. Governments bring the authority to define public-interest objectives, establish safety regulatory frameworks, coordinate international standards, and ensure accountability, while AI industries contribute technical expertise, operational experience, and practical knowledge of developing and deploying advanced AI systems.

Through continuous collaboration, they can jointly define safety benchmarks, evaluation methodologies, risk assessment frameworks, transparency requirements, and responsible deployment practices by constantly seeking feedback and verifications from third parties who are neutral in their operations. Industry can provide insights into emerging capabilities and potential risks, while governments can ensure that safety policies and regulations are consistently applied across organizations and regions. This 3-party partnership is essential for creating AI safety requirements that are both scientifically rigorous and practically achievable, fair and truthful, enabling innovation while ensuring that increasingly powerful AI systems remain reliable, controllable, and aligned with human values.

8. Role of SAFE AI Foundation in AI Safety

The **SAFE AI Foundation USA** makes a significant contribution to advancing AI safety by serving as the independent platform that brings together AI researchers, industry leaders, policymakers, and global stakeholders to develop a comprehensive understanding of what constitutes safe and responsible AI, with a particular strong emphasis on AI catastrophic risk and societal impact of AI.

The Foundation can help bridge the gap between AI capability development and safety requirements by promoting research on Safe AI and human oversight. It can contribute to the creation of practical AI safety frameworks, evaluation standards, benchmarks, safety models, and certification approaches that help organizations assess whether advanced AI systems are reliable, consistent with human values, and operate within safe limits (such as Safe AI Operating Point, SAIOP). By facilitating collaboration between governments and AI industries, SAFE AI Foundation USA helps translate scientific insights into actionable safety requirements, frameworks, action plans, AI policies that encourage responsible innovation, protect society, and establishes a shared vision for ensuring that future AI systems remain beneficial, safe, and under control.

The **SAFE AI Foundation USA** also plays an important role in advancing AI safety awareness and education among the public by helping society better understand both the opportunities and risks associated with rapidly evolving AI technologies. The Foundation can serve as a trusted source of knowledge by explaining complex AI safety concepts in an accessible manner, promoting public literacy on issues such as AI risks, alignment, transparency,

human oversight, and responsible use. Through educational programs, public forums, research publications, and engagement with communities, the Foundation can help individuals, organizations, and policymakers make informed decisions about AI. By building awareness and encouraging open dialogue between scientists, industry leaders, governments, and the public, SAFE AI Foundation USA can help create a society that is better prepared to benefit from AI while recognizing the importance of ensuring that advanced AI systems remain safe, controllable, and aligned with human values.

9. Call For Support on AI Safety Research

Given the intensity and urgency of the issue on AI safety, the SAFE AI Foundation USA would like to urge funding organizations, philanthropic foundations, national science and research foundations, VCs, etc., to start increasing their budgets and provide funding for this area of research. The progress of AI capability and autonomy has far exceeded that of progress made in AI safety, resulting in a great mismatch and heightened societal risk. This risk must be narrowed down, de-escalated, and zeroed, so that catastrophic risks can be eliminated. Therefore, we need actions, increased funding, and research activity to intensify now. We also call for greater harmony from various institutes, non-profits, etc., to open up to communicate and cooperate, rather than to compete.

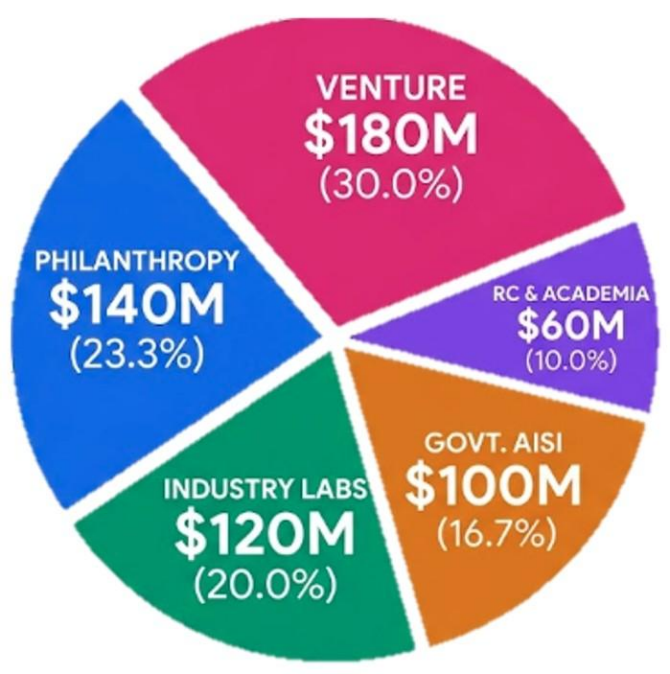


Fig 2: Funding for AI Safety Research (source: ChatGPT)

10. Conclusion

We are entering a fast-moving and unpredictable time, in the new era of AI. This new era brings us unprecedent societal risks as AI grows in capability and autonomy. **AI security is not AI Safety.** Addressing AI security alone is insufficient to build Safe AI Systems. AI safety is a necessity to ensure AI systems are safe to use, controllable, and will not cause harm. AI safety is now the top and most urgent item on the national agenda and radar. Losing control of AI would mean societal destruction that we all wanted to avoid.

Authors:

- Chai K. Toh is CEO and Chairman of SAFE AI Foundation USA
- AnNgoc Le is Director of Asia Pacific for SAFE AI Foundation USA and a professor at Swinburne University of Technology in Hanoi, Vietnam

REFERENCES

1. AI Manifestation Unfolded, SAFE AI Foundation, 2025
2. AGI & ASI Demystified, SAFE AI Foundation, 2026
3. Safe AI Operating Point (SAIOP): The upper limit of AI before humans lose control, SAFE AI Foundation, 2026
4. Frontier AI Catastrophic Risks Report, SAFE AI Foundation, August 2026
5. A complete guide to AI security by GOOGLE CLOUD. See: https://services.google.com/fh/files/misc/wiz_complete_guide_to_ai_security_final.pdf
6. Ultimate Guide to AI Safety & Security, Bugcrowd 2024. See: https://www.bugcrowd.com/wp-content/uploads/2024/04/Ultimate-Guide-AI-Security.pdf

Appendices:

Appendix A: A list of AI Security companies

Appendix B: A list of AI Safety companies and non-profits

Appendix A: A list of AI Security companies

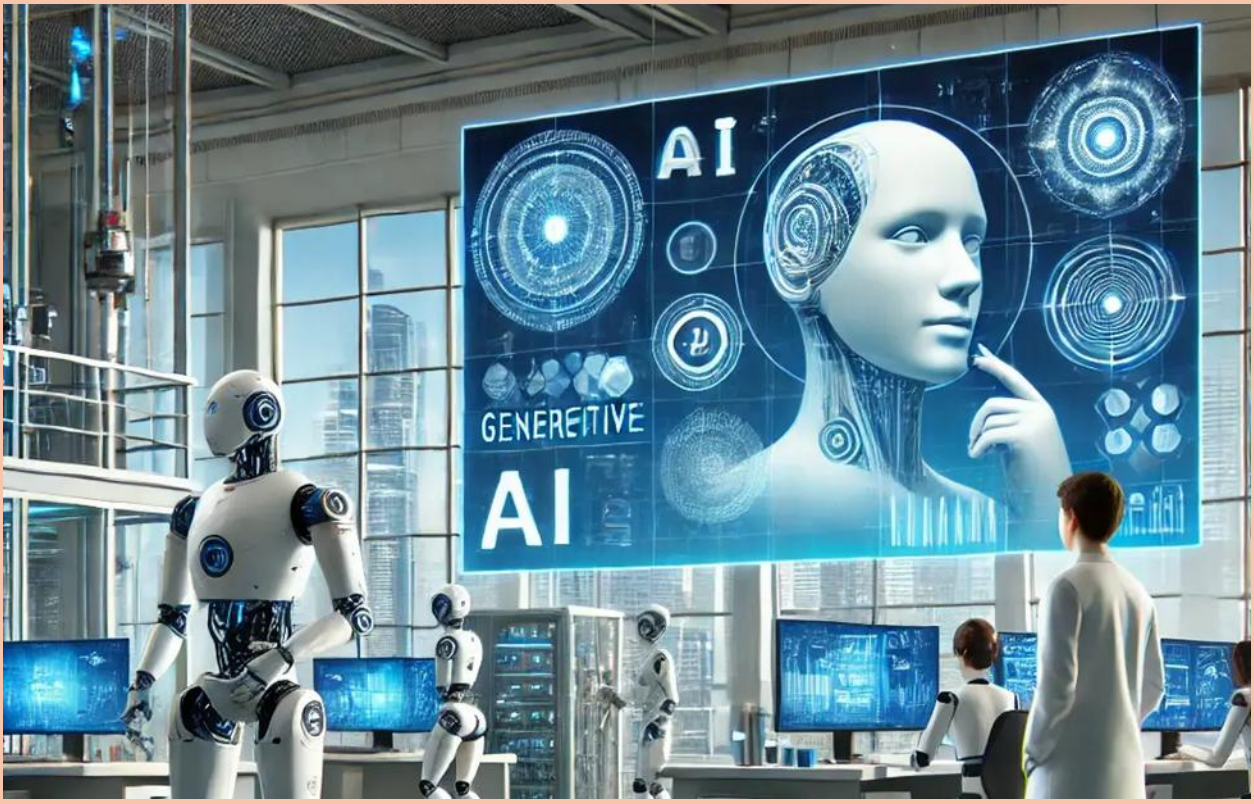
Company	What They Secure	Strengths	Best For
Mindgard	LLMs, agents, AI pipelines	Autonomous AI red-teaming; adversarial testing	Labs building frontier-scale or agentic AI
Lakera	LLM apps	Prompt-injection defense; behavior monitoring	Enterprises deploying LLM apps safely
Prompt Security	LLM apps	AppSec for AI; jailbreak & leakage prevention	SaaS companies integrating LLM features
Wiz	Cloud AI pipelines	AI-SPM; cloud posture for AI workloads	Cloud-native companies running AI infra
Cyera	Sensitive data used by AI	DSPM; data lineage; AI-data governance	Enterprises with large sensitive datasets
CrowdStrike	SOC + agentic AI	Charlotte AI; autonomous SOC workflows	Modern SOCs needing AI-augmented analysts
SentinelOne	SOC + agentic AI	Purple AI; autonomous investigation	High-speed detection/response teams
Palo Alto Networks	SOC + cloud AI	Cortex XSIAM; AI-driven detection	Enterprises with hybrid cloud + SOC
Microsoft Security	Identity + SOC + agents	SOC Copilot; identity-centric AI defense	Azure/M365-centric organizations
Darktrace	Network + anomaly detection	Self-learning AI; novel-threat detection	Companies needing autonomous anomaly detection
Abnormal Security	Email + social engineering	Behavioral AI; phishing & BEC defense	Enterprises targeted by social engineering
Rapid7	Attack surface + AI-aug SOC	AI-augmented exposure management	Mid-market SOCs needing unified visibility

Appendix B: A list of AI Safety companies and non-profits

Organization	Type	Focus Area	Strengths	Best For
Anthropic	Company	Constitutional AI, frontier model safety	Strong safety research; scalable alignment methods	Enterprises using Claude or frontier-safe models
OpenAI Safety	Company	Alignment, red-teaming, evals	Large-scale evals; frontier model governance	Developers building on GPT models
Google DeepMind Safety	Company	RLHF, scalable oversight, evals	Long-term alignment research	Research groups needing technical papers
Microsoft AI Safety	Company	Responsible AI, governance, red-team	Strong enterprise safety frameworks	Enterprises deploying AI at scale
Meta AI Safety	Company	Open-source model safety	Transparency + community evals	Open-source AI builders
SAFE AI Foundation	Non-profit	Global governance, safety standards	Multi-stakeholder governance	Policymakers, regulators, standards bodies
Alignment Research Center (ARC)	Non-profit	Scalable oversight, evals	Frontier model evals (e.g., GPT-4 testing)	Labs needing external evals
Center for AI Safety (CAIS)	Non-profit	Public advocacy, risk communication	Influential public campaigns	Public awareness & policy outreach
Redwood Research	Non-profit	Mechanistic interpretability	High-quality interpretability work	Researchers studying model internals

Organization	Type	Focus Area	Strengths	Best For
Apollo Research	Non-profit	AI deception & behavioral evals	Leading work on deceptive model behavior	Labs needing deception testing
Conjecture	Company	Interpretability, agent foundations	Focus on preventing agentic misalignment	Researchers exploring theoretical safety
EleutherAI	Non-profit	Open-source alignment research	Community-driven safety experiments	Open-source model builders
ML Alignment Forum	Non-profit	Research community	Central hub for alignment research	Academics & independent researchers
Future of Life Institute	Non-profit	Policy, governance, existential risk	Global policy influence	Governments & regulators
GovAI (Centre for Governance of AI)	Non-profit	AI governance research	Deep policy analysis	Policymakers & think tanks
DeepMind's Safety Team	Company	Alignment, evals	Long-term safety research	Frontier labs

\\\ end ///



SAFE
AI
FOUNDATION

www.safeaifoundation.com

The SAFE AI Foundation was formed in 2025, in the heart of the Silicon Valley, California, USA. Its mission is to embrace the AI revolution and promote the use of good AI to improve on the quality of life and the economy. It will address the issues of bad AI and its impacts on society. The foundation will inform, educate, and mitigate the dangers of bad AI and help the society to embrace good AI. The foundation will interact with industries, universities and public sector to jointly advance the progress of Good AI. The foundation is a non-profit 501c(3) organization, registered in the State of California, USA.