GRADIENT INSTITUTE

2025

Impact Report

Helping Australians understand, steer and respond to AI as it transforms our world

In 2025, Gradient Institute shaped Australia's national AI guidance, contributed to global AI safety science, and helped hundreds of organisations build the capabilities to develop, deploy and use trustworthy AI systems.

gradientinstitute.org
ACNC Registered Charity | ABN 29 631 761 469

CONTENTS

ABOUT GRADIENT INSTITUTE

A nonprofit research organisation that values humanity, rigour and independence.

Gradient Institute was formed in 2019 to ensure that the decisions shaping Artificial Intelligence's development and use are made with the clearest possible understanding of what is actually at stake. We do not sell AI products. We have no shareholders. We conduct, distill, and interpret scientific research so that leaders, communities, and the public can make well-informed decisions on AI and help Australia benefit from the potential of AI while avoiding many of the harms.

Our Theory of Change

The gap for good decision-making

AI systems have become incredibly powerful and impactful. There is a gap between what AI systems can do and what the people making decisions about their use genuinely understand of them, and this gap continues to widen.

Incentive structures, political pressures and competition often push decision-makers towards quick decisions. A key foundation for good judgment - human-centred, independent, rigorously grounded analysis - is often missing when important decisions are made. We believe that this will inevitably lead to negative outcomes for individuals, organisations and society.

Providing clarity and understanding

We believe that better understanding is a necessary foundation for policy makers, regulators, organisations and individuals to make good decisions about AI. We build better understanding through conducting our own research, making frontier research accessible to decision-makers, delivering education and training, and contributing to public discourse.

Advocating for systems change

We believe that good decisions about AI can be further enabled by designing societal and organisational systems that require or incentivise decision-makers to consider the science of AI and implications for how it is built, used and governed. We advocate for systems change by working with governments, industry and researchers to design governance structures and practical guidance and tools that can manage uncertainty, and through targeted advisory work and policy submissions.

The AI transformation benefits everyone

Decision-makers are more likely to make decisions which help rather than harm society when they are empowered with: (i) a strong understanding of AI informed by the best scientific evidence available and (ii) the appropriate societal and organisational systems around them. With this, individuals, organisations and society can benefit from the enormous and rapidly accelerating potential of AI systems, without being exposed to disproportionate risks.

A LETTER FROM THE CEO

Whether AI progress in 2025 was genuinely exponential is still debatable. It certainly felt that way. And the institutions charged with governing it responded accordingly.

Governments, boardrooms and regulatory agencies saw a great deal of AI activity. National AI strategies were published. Chief AI Officers were appointed. Consulting firms were commissioned. Deals were signed with frontier model providers. Educational institutions debated how to teach. Media companies pumped out stories. Citizens tried to make sense of what it all meant for their jobs, their privacy, their futures. By most visible measures, the institutions that govern our economy and public life are taking AI seriously.

The harder question is to what extent that activity is matched by the understanding needed to appropriately govern this general purpose technology, whose consequences we still can't fathom. The gap between what AI systems can do and what the people making decisions about them genuinely understand seems to be widening.

Gradient Institute was built for this problem: the harder, slower work of building genuine understanding, both among the people whose decisions shape how this technology develops and among the public whose lives it will shape. We do this through rigorous independent analysis, at a deliberate distance from lobbying and the competing interests of the industry itself.

In 2025, that work produced outcomes that would have been difficult to imagine just two years ago: official government guidance for all businesses, internationally impactful research, influence on government policy and new science-based AI literacy programs. Together, they demonstrate that the gap between AI capability and its understanding by decision-makers can be narrowed when the work is done with patience, care, and a long-term view.

All of this reflects years of investment by a small team in the credibility, relationships and rigour that serious influence requires. It also reflects the commitment of funders and partners who share our conviction that humanity, rigour and independence are essential foundations for this work.

Thank you for making that possible.

Bill Simpson-Young

Bill Simpson-Young
Co-founder and Chief Executive, Gradient Institute

2025 BY THE NUMBERS

POLICY INFLUENCE

80%
Australian Government AI guides for industry shaped by Gradient[1]

12
Advisory boards and working groups

4
Government submissions

RESEARCH & KNOWLEDGE

1st
World-first report on risk methodology for multi-agent AI systems

3
Testing exercises with the International Network of AI Safety Institutes

6
Editions of our horizon-scanning report made available for partner orgs

EDUCATION & OUTREACH

45
Education sessions delivered

4,500+
People reached by talks and training, workshops and training courses

140+
Not-for-profits reached through education programs

27,000+
Interactions with AI guidance materials

PARTNERSHIPS

40+
Organisations directly engaged with

8
International networks participated in

15
Startups advised on responsible AI

[1] This figure shows the percentage of AI guidelines for industry published by the Australian Government's Department of Industry, Science and Resources from 2019 to 2025 that were co-written or influenced by Gradient Institute. Publications are available from https://www.industry.gov.au/publications.

STORIES OF IMPACT

Five stories from 2025 that show how Gradient's work created real-world impact.

Equipping Australian businesses to adopt AI safely

Australian businesses - particularly small and medium sized ones - were seeking accessible, actionable guidance on deploying AI safely, yet improvisation was often the only option.

Recognising the gap, the Australian Government's National AI Centre (NAIC) engaged Gradient in July 2025 to co-develop practical guidance for the safe and responsible use of AI suitable for Australia-wide use. Gradient and NAIC created six essential practices for AI governance that are pragmatic, actionable, and designed for organisations with real constraints, alongside an AI policy template, an AI screening tool, and an AI register template.[2] The work drew on Gradient's engagements with not-for-profits, and carefully considered how to support organisations to manage the uncertainty inherent in AI risks and opportunities.

Guidance for AI Adoption: Foundations

AI policy guide and template

AI screening tool

AI register template

[2] Gradient developed the 6 essential practices, the Foundations document and the templates with NAIC. The Implementation Practices document was developed by NAIC based on the Voluntary AI Safety Standard with involvement from CSIRO's Data61, the Human Technology Institute at the University of Technology Sydney and Gradient Institute. They are available at: https://www.industry.gov.au/publications/guidance-for-ai-adoption

In October, the Australian Government released the Guidance for AI Adoption at the Committee for Economic Development of Australia (CEDA) and NAIC AI Leadership Summit. The framework has since been adopted as the official advice on AI at business.gov.au, embedded in the Future Skills Organisation's national AI training, and taken up by TAFE in its teaching. Many organisations across Australia have adopted the practices and templates directly.

> **WHY IT MATTERS:** Rather than guiding one organisation at a time, Gradient shaped the standard every Australian business can start from, by developing practical, government-backed advice in under six months.

"The Guidance is a call to arms for every Australian business large and small."

Assistant Minister Andrew Charlton[3]

Delivering a world-first report on multi-agent AI risk

As AI agents begin working together in complex multi-agent systems, their interactions create new safety risks that are difficult to anticipate and govern. Until recently, there was no publicly available methodology designed specifically for analysing the risks of multiple interacting agents within an organisation.

Funded by the Department of Industry, Science and Resources (DISR) as part of Australia's contribution to international AI safety collaboration, Gradient produced *Risk Analysis Methodology for Governed LLM-based Multi-Agent Systems*, the first report globally to focus specifically on risk methodologies for multi-agent AI systems under single-organisation governance.

[3] CEDA AI Leadership Summit, Brisbane, October 2025.

Before publication, drafts were reviewed by staff from the UK AI Security Institute, CSIRO's Data61, the University of Sydney, Australian National University (ANU), Commonwealth Bank of Australia, IAG, Suncorp and the NSW Government, whose feedback helped sharpen the methodology and ground it in real-world use cases and emerging risks.

Published on the arXiv website for scientific papers in August 2025, the report has since been used to implement governance checklists at major Australian organisations.[4] It is referenced in the International AI Safety Report (February 2026), the leading global synthesis of the field, and cited extensively by the University of California, Berkeley's Center for Long-Term Cybersecurity in their work extending the US Government's AI Risk Management Framework to manage the risks of AI agents.

> **WHY IT MATTERS:** Governing a technology is impossible without the tools to assess it. Gradient built the first such tool for multi-agent AI risk within a governed environment, which is now informing practice at Australian organisations, informing researchers internationally and well-positioned to influence standards efforts.

[4] The report is available at: https://arxiv.org/abs/2508.05687

OUR RESEARCH PROGRAM: PURSUING CRITICAL QUESTIONS IN AI SAFETY

The multi-agent report is part of Gradient's broader research program. During 2025, we also explored the following questions.

Can we detect what an AI system can do before those capabilities appear in use? With the UK AI Security Institute and AI safety research organisation Timaeus, Gradient worked on methods to identify what AI models are capable of before those abilities become visible. This matters because models sometimes have capabilities their developers haven't tested for or don't yet know about. The UK AI Security Institute reviewed the work and encouraged its publication which will happen in 2026.

How can we evaluate the cybersecurity capabilities of AI models? Through the International Network of AI Safety Institutes, Gradient helped develop methodologies to measure the cybersecurity capabilities of large AI models. Partners included CSIRO's Data61, DISR, and the UK AI Security Institute.

What do scientists worldwide currently understand about AI risks and safety? Gradient was invited to contribute to and review the February 2026 edition of the International AI Safety Report, the leading worldwide synthesis of AI risks and safety research. Gradient reviewed specific technical components, provided substantial input and reviewed multiple full drafts. It is the only Australian technical civil society organisation acknowledged as a contributor in the report. As mentioned above, Gradient also has its multi-agent research referenced in the report.

How can industry and government adapt to AI? Gradient is a partner on two Australian Research Council (ARC) Linkage grants applying research to real-world governance challenges:

- *How can insurance be made socially responsible in the age of AI?* with ANU, the University of Sydney, and the insurer IAG.
- *How can policy adapt to keep pace with rapidly evolving AI?* with RMIT and other partners.

Gradient is also helping develop Australia's next generation of AI safety researchers with ANU and co-funding from the Australian Government's Next Generation Graduates Program. Student work in 2025 included research on how AI models learn during an early stage of training - a period that shapes what they can and can't do later - and how systems of multiple interacting AI agents can be made safe and reliable.

Building AI capability across Australia's not-for-profits

Australia's not-for-profits work closely with vulnerable and marginalised communities and have significant expertise in stakeholder engagement, impact assessment, and risk management. To put those strengths to use with AI, even as they operate under tight funding and often limited in-house technical capacity, the sector needs accessible tools and technical support tailored to its context.

With a grant from Google.org, Google's philanthropic arm, Gradient worked to build responsible AI capability across Australia's not-for-profit sector. Eight educational webinars were attended by more than 350 attendees from over 140 not-for-profits. Three eLearning modules extended the reach through national not-for-profit platforms. Gradient also developed open-source governance templates and provided hands-on support to three not-for-profits to trial the templates before wider release.

The templates developed through this work formed the foundations of some of the templates later adopted by the NAIC (see earlier story), carrying the sector's commitment to keeping people at the centre into the national guidance for all Australian businesses. Infoxchange have since developed a course for not-for-profits building on this work.[5] Many not-for-profits are now better equipped to use AI responsibly and freely available resources continue to enable not-for-profits to build their AI capability.

> WHY IT MATTERS: When not-for-profits use AI well, the communities most often overlooked by new technologies can benefit from them. Gradient strengthened the sector by providing tools and resources for not-for-profits to continue or start exploring opportunities to increase their impact using AI.

[5] The course is accessible at: https://learning.infoxchange.org/enrol/index.php?id=412

Making the case for an AI Safety Institute

Until recently, Australia had no dedicated body to research, monitor, or measure the capabilities and risks of AI, and no sovereign scientific capability to guide its response.

Since July 2023, Gradient made a sustained case for change:

- four formal submissions to the Australian Treasury, the DISR, and the Productivity Commission;
- a co-authored paper with Good Ancestors Policy; and
- a joint in-person presentation to DISR in Canberra.

In November 2025, the Australian Government announced a new AI Safety Institute, funded at $30 million over four years, with a scope closely aligned to what Gradient had proposed. Many organisations and partners contributed to this positive outcome. Gradient had made the case formally and repeatedly since July 2023.

> **WHY IT MATTERS:** A country's capacity to govern AI safely depends on sovereign scientific and technical capability. Gradient helped make the case for building that capability in Australia, and the government acted.

Collaborating with international organisations on technical governance

AI standards and codes of practice are being written now, setting the technical specifications that will define what counts as safe and responsible AI for regulators, auditors, and courts. Countries without experts at those drafting tables inherit rules shaped by others' priorities.

Throughout 2025, Gradient staff brought Australian technical expertise into the rooms where those rules are being written. Staff contributed to ISO/IEC standards development, participated in the EU's General-Purpose AI Code of Practice through its risk assessment and technical risk mitigation working groups, and attended global standardisation meetings and plenaries.

Gradient was invited to join the International Association for Safe and Ethical AI as an affiliate organisation, and contributed to joint testing exercises with the International Network of AI Safety Institutes, developing novel cybersecurity evaluation methodologies in collaboration with CSIRO's Data61. To strengthen the relationships which keep Gradient's work connected to the global frontier, Gradient also met with teams such as those at Google DeepMind, Control AI, and the UK AI Security Institute.

Gradient helped shape the technical standards, codes, and scientific assessments that will govern AI globally. Gradient staff were able to contribute perspectives which were independent of commercial or political agendas, to be considered alongside the perspectives of some of the world's largest technology companies and regulatory bodies.

> WHY IT MATTERS: Industry has significant influence over the technical rules being written for AI globally. Gradient ensures Australian public-interest expertise is represented too.

LOOKING AHEAD: 2026

AI is accelerating. The decisions being made now about regulation, safety standards, institutional capability, and public understanding will shape whether that acceleration serves people or harms them. Gradient enters 2026 with deeper influence, stronger partnerships, and a clearer mandate than at any point in our 7-year history.

Equipping AI businesses to adopt AI safely

We are finding a growing need for organisations' assurance of the safe and responsible operation of their AI agent systems and will be growing work in this area given its importance for Australian society and the lack of effective approaches for doing this.

Taking our multi-agent research further

We are preparing for major research projects to expand our work on multi-agent risks to agents operating across organisations and simulation platforms for agent safety testing.

Building AI capability across the Australian public and decision-makers

Under our new science-based AI literacy initiative launched in January 2026, funded by Google.org, we will be building educational events and resources, and cross-sector dialogue to ground Australia's AI discourse in evidence rather than myths.

Supporting the Australian AI Safety Institute

With the establishment of Australia's AI Safety Institute in the first half of 2026, we hope to be able to collaborate with them in the identification and understanding of emerging AI harms and help Australia respond to these quickly and effectively.

Collaborating with international organisations

We plan to further build our international collaborations on research and standardisation, particularly on AI agent governance and risk management, as the opportunities and risks of AI are global and it will take a global perspective to steer the AI future to be a positive one.

Preparing for an AI future

We expect that 2026 will be a defining year in the development of AI capabilities, use and impacts. In particular, the improved capabilities of AI agents will lead to different ways of working across many organisational roles. We plan to do research on new societal and organisational systems that will help this happen in a way that improves the lives of humans, rather than harming them.

OUR PARTNERS & COLLABORATORS

Most of the work we do, we do with others. Gradient's impact is amplified through sustained partnerships with organisations that share our commitment to a humane, independent, and rigorous approach to AI. We would like to recognise the following partners and collaborators.

Government Collaborators

We worked with the Australian Government's Department of Industry, Science and Resources (DISR) including its National AI Centre (NAIC) to provide guidance, education and research on safe and responsible AI for Australian organisations. We also worked with other federal and state government agencies (including the Office of AI in the NSW Government) to improve these capabilities within government itself. We provided technical guidance through discussions, education and submissions to aid in sound AI policy development that supports a flourishing Australian society.

International Collaborators

AI is a global phenomenon and Australian society benefits from global expertise. We worked with the UK AI Security Institute, International Network of AI Safety Institutes, EU AI Code of Practice working groups, International Association for Safe and Ethical AI, the team developing the International Scientific Report on AI Safety, International Standards Organisation and the Timaeus AI safety scientific research organisation.

Research Collaborators

We collaborated with researchers from the University of Sydney, ANU, CSIRO's Data61, Timaeus, and the International Scientific Report on AI Safety review team.

Expert and Advisory Boards

We contributed to AI policy and practice through expert and advisory boards including the: NSW Government's AI Review Committee, EU AI Code of Practice (two working groups), Standards Australia AI committees, Paul Ramsay Foundation AI Expert Panel, Human Technology Institute AI Policy & Regulation Expert Reference Group, Human Technology Institute Thrive Expert Reference Group, Financial Counselling Industry Fund Innovation Advisory Panel, the ANU Computing Advisory Board, the Technical Committee for Specialist AI project at Future Skills Organisation and others.

Civil Society Collaborators

We worked with many not-for-profits through our AI capability uplift program as described above. We also collaborated with other civil society organisations during the year including Good Ancestors, the Sydney AI Safety Space, the Human Technology Institute at UTS and the Tech Policy Design Institute.

Industry and Advisory

We advised and worked with large-scale deployers of AI systems to support their efforts to use AI safely and responsibly for the benefit of Australian citizens. This included Australia's largest general insurer IAG and Australia's largest bank Commonwealth Bank of Australia. With Gradient's assistance since 2021, the latter was judged in 2025 to be the second highest bank globally in the responsible use of AI.

Foundation Members

IAG and the University of Sydney serve as foundation members. IAG has provided significant financial grant support since Gradient Institute's founding, and this commitment underpins our capacity to conduct research into safe and responsible AI independently.

AI's future is not inevitable.
It is being *chosen*.

We exist to bring clarity to the choosing.

To discuss how your support can help Gradient Institute extend this work:
info@gradientinstitute.org

GRADIENT INSTITUTE
Sydney Knowledge Hub, Merewether Building H04
University of Sydney NSW 2006, Australia
gradientinstitute.org