Shaping the Future Together: AI, Innovation, Inclusion and Resilience in a Fractured World

A Report by Gates Cambridge Scholars

Introduction

From April 29 to May 1, Bridgehouse, the Gates Cambridge Trust, and the Intellectual Forum at Jesus College, Cambridge, organised an intellectual retreat in Cambridge for senior executives and CEOs. The delegates engaged with entrepreneurs and academics, joining seminars centred on four distinct topics: AI and Human Judgement in Decision-Making, Innovation and Inclusion in Technological Systems, Resilience in a Polycrisis Era, and Public Trust in the Age of Deepfakes and Polarisation.

The session on **AI and Human Judgement in Decision-Making** was led by Professor Mateja Jamnik and Gates Cambridge Scholars Hannah Claus and Meiru Zhang. The discussion engaged with questions of cognitive identity, AI functionality, the representativeness of AI models, and the role of human beings in ensuring AI accuracy.

Led by Professor Bryan Zhang and Gates Cambridge Scholars Abdullah Hasan Safir and Aliva Salmeen, the **Innovation and Inclusion in Technological Systems** session explored how contemporary AI and technological systems are reshaping our world in ways that are deeply entangled with questions of power, representation, and global inequality. Beginning with the need for the financial system to adapt swiftly to ensure that innovation is balanced with consumer protection, market integrity, and financial stability, the session also drew on critical scholarship, ethnographic fieldwork, and community-centred design practices to argue that innovation cannot be understood purely as technical progress. Rather, it must also be evaluated through the lens of inclusion, justice, and whose knowledge systems are allowed to shape the future of AI.

The **Resilience in a Polycrisis Era** session was led by Professor Charlotte Hammer and Gates Cambridge Scholars Albert Van Wijngaarden and Nikita Jha. Across the discussion, spanning polypandemics, organisational antifragility, and polar climate interventions, the session drew on global, interdisciplinary research to offer diverse perspectives on a central question: what does it take for systems, organisations, and countries to hold together under compounding, unprecedented pressures in an increasingly complex world?

Led by Professor Andreas Vlachos and Gates Cambridge Scholar Alex Mentzel, the **Public Trust in the Age of Deepfakes and Polarisation** session examined the vulnerabilities of our shared information ecosystem. It analysed the historical continuity of propaganda alongside the distinct challenges posed by the collapsed costs of generating modern misinformation, taking a philosophy-informed approach to exploring how institutions and verification networks must evolve to preserve corporate and social stability.

Across these seminars, discussions coalesced around three key themes: trust, relevance, and opportunities. Discussing each of these in turn, the report distils the key takeaways from the sessions.

Trust

Trust emerged as one of the defining concerns of the seminar, central to the successful navigation of global uncertainties and challenges. At the level of the human-AI interface, delegates across sectors expressed significant anxiety about whether AI systems can be relied upon to deliver consistent, accurate outputs. A striking illustration came from a delegate who described consulting an AI tool about government policy on two separate occasions using a major frontier model, only to receive entirely different conclusions. This unpredictability, rooted in probabilistic sampling rather than error, unsettled confidence in AI as an advisory resource. The consensus response was practical: use multiple AI tools to cross-check answers, demand verifiable reasoning chains, and resist allowing emotional framing to colour queries.

This internal unpredictability is further complicated by the question of expertise and domain misalignment. Delegates observed that the gap between professional and non-expert use of AI is narrowing rapidly as tools become easier to access, yet the ability to critically evaluate AI outputs has not kept pace. In high-stakes domains, such as banking, legal advice, and public policy, organisations are proceeding cautiously, with agentic AI largely held back due to concerns about faithfulness and reliability. The practical takeaway here is clear: human oversight remains non-negotiable in consequential decisions, and organisations should be explicit about where that threshold lies.

A further dimension of trust that the seminar flagged, but did not explore in full depth, is multi-agent safety. As AI moves from single-agent prototypes to orchestrated multi-agent pipelines, new risks emerge: collision between agents acting on conflicting objectives, runaway feedback loops, and prompt-injection chains where one agent manipulates another. These threats sit outside the scope of conventional line management, supply chain, and cybersecurity frameworks, which were not designed for agent-to-agent interactions. Establishing trust in agentic AI will require not only verifying individual model outputs, but verifying the behaviour of the system as a whole.

Yet, technical verification is only part of the challenge. The integrity of the information ecosystem itself is at stake. One session opened with Jonathan Swift's line from 1710 — ‘Falsehood flies, and truth comes limping after it', prompting the question: was it true then, and is it true now? Swift wrote it in *The Examiner*, a partisan Tory paper, in the thick of one of England's first organised propaganda wars, with Whig and Tory writers — Addison and Steele on one side, Swift and Defoe on the other — firing pamphlets across London to shape public opinion. So the line was almost certainly true for Swift, but it was also itself a partisan statement. As for today, the best-known evidence is Vosoughi, Roy and Aral's 2018 study in *Science*, which traced how far and how widely news stories spread on Twitter and found that stories rated false by fact-checkers travelled faster and reached more people than true ones — Swift's intuition holding on at least one major modern platform. The picture is not uniform, though: González-Bailón and colleagues, studying Facebook during the 2020 US election, found misinformation spreading more slowly, and the Twitter study rests on a single platform that predates generative AI. So this is not a new worry. It was shared by Thucydides in the fifth century BCE and by Montaigne in the sixteenth century CE. And each previous information-technology wave eventually produced an institutional adjustment: the printing press gave us libel law and editorial standards, broadcast gave us regulators, and the internet gave us professional fact-checking organisations such as Full Fact and Africa Check.

It might be said, then, that what is genuinely new about the generative-AI moment is that the cost of producing, distributing, and verifying information has collapsed all at once. A practical question for delegates follows from this fact: what does your organisation's institutional adjustment look like, and who checks what a model says before it reaches a customer or a decision-maker? The binding constraint for leaders is rarely access to data, or even to verified facts — it is whether people feel they can act on what they are seeing. The leadership implication, for public institutions and private firms alike, is that in a post-AI environment it is no longer enough for a source to be reputable. People need to see the process, and to feel they were part of it rather than dictated to by it.

This need for inclusion and transparency extends from organisational workflows to the global stage, where trust is compromised by systemic inequities. Discussions highlighted growing critiques of large technology companies for enabling commercial surveillance, opacity, algorithmic bias, and asymmetrical concentrations of power. These concerns are particularly significant for Majority World communities, whose experiences and realities are often marginalised in the design and governance of AI systems. The presentation connected these issues with the concept of *epistemic injustice*, emphasising how people can be excluded not only economically or politically, but also as legitimate knowers and contributors to technological futures.

Conversations around AI and preventive healthcare further emphasised that trust depends not only on technical performance, but also on whether systems are equitable, clinically meaningful, and appropriate for local populations. Risk prediction tools demonstrate how AI can support earlier identification of individuals at elevated risk of disease and potentially enable a shift from reactive healthcare toward more proactive prevention. However, a Gates Cambridge Scholar stressed that 'one model does not fit all'. Models developed using data from one population may perform poorly in another due to differences in demographics, healthcare systems, disease patterns, and social determinants of health. Local validation and recalibration therefore emerged as essential rather than optional steps, particularly for low- and middle-income countries. Through further examples from AI ethics, image generation, and refugee and slum communities in Bangladesh, the seminar argued that trust in AI systems depends on whether technological infrastructures genuinely recognise and empower marginalised voices rather than merely extracting data from them.

When scaled to the geopolitical level, these fractures reveal the immense macroeconomic costs of misplaced trust. Just as citizens must trust systems, nations must navigate their own dependencies; indeed, the COVID-19 pandemic revealed the costs of over-reliance on single countries for critical resources and supply chains, which left governments exposed when those relationships failed. While the pandemic demonstrated the dangers of concentrated dependency, discussions acknowledged how some degree of interdependence between countries is a core characteristic of an interconnected world. No country is entirely self-sufficient, and the international order has always functioned on the assumption that countries will rely on each other. The more precise concern is the kind of interdependence being built, and whether it rests on a foundation stable enough to hold under stress. Trust and cooperation are central to the robustness of this foundation, and this foundation, in turn, is central to systemic antifragility. Systems that can absorb and evolve from shocks are often those embedded in relationship networks characterised by open information flows, upheld commitments, and acknowledgement of failures. Without those conditions, the network fails to buffer against risk.

Delegates raised questions around whether placing such emphasis on trust and shared values was idealistic given the present state of geopolitics. With the previous world order currently in a state of flux, countries are not, in practice, operating on a basis of mutual trust. The response from the session was to highlight that the post-war international order was built as much on

pragmatism as ethical consensus. Countries chose to negotiate common frameworks and sit at shared tables because war is expensive, resource-intensive, and ultimately self-defeating. The case for cooperation is a self-interested one, and thereby perhaps more sustainable than reliance on trust alone.

Relevance

A recurring undercurrent throughout the retreat was the question of what remains distinctively human and valuable as AI capabilities expand. Some seminars explored whether large language models possess genuine intelligence, with the scholars proposing that inference alone does not constitute intelligence, and noting that a universally agreed-upon definition of intelligence remains elusive. This perspective aligns with the work of philosopher of technology Luciano Floridi, who demonstrates that large language models do not ground meaning in the world themselves but instead inherit it from human language and practice. Delegates converged on the view that what humans retain, and what AI currently lacks, is initiative: the capacity to form goals, to want something, to act from an internal sense of purpose rather than in response to a prompt. Creativity and wisdom were similarly identified as human qualities that go beyond the pattern recognition and synthesis at which LLMs excel.

Crucially, empirical research reframes this distinctively human judgement as collective rather than individual. Asked to identify AI-generated 'deepfake' text, expert readers performed no better than chance alone — but after group discussion they identified the fakes far more accurately. The same pattern appears in the classic Wason reasoning task, where individuals succeed 10–20% of the time and small groups around 80%. Human relevance, then, lies less in lone expertise than in our evolved capacity to reason together, with the most productive groups being those that probe each other for reasons and tolerate a diversity of initial answers.

This collective human relevance is as much about the present capabilities and limits of AI as about where it is heading, which raises pointed questions about education and the future of work. Current empirical evidence shows that agent performance roughly halves when measured on end-to-end tasks rather than isolated subtasks, suggesting that the gap between completing a single step and completing an entire job remains large. For now, the human role in orchestrating, sequencing, and quality-checking multi-step work is far from obsolete. However, knowledge and expertise, once a reliable form of professional currency, are being devalued. This creates pressure on individuals and institutions alike to articulate what human judgement uniquely contributes, prompting reflection about what skills will matter for the next generation if routine decision-making is increasingly delegated to AI, and whether people without the ability to craft effective prompts or evaluate AI outputs will be effectively excluded from the benefits. The pressing issue is how quickly that gap closes, and whether individuals and organisations are building the judgement skills to stay ahead of it.

The question of relevance of human perspectives in an increasingly AI-driven world applies not just to the development of AI but also to companies, institutions, and countries. Societies and organisations that ignore cultural plurality, local knowledge, and community participation risk becoming socially and politically disconnected from the people they claim to serve. AI systems built primarily around Western assumptions and infrastructures often fail to meaningfully represent diverse cultures, languages, and artistic practices.

In response to these dynamics, governments, companies, and research institutions must remain relevant by investing in inclusive innovation practices, engaging with communities directly, and recognising diverse epistemologies as essential to technological progress. Future leadership in AI will not come only from computational scale, but also from the ability to design systems that are socially grounded, culturally aware, globally inclusive, and responsive to the lived realities of diverse communities.

When scaled up to organisational and geopolitical levels, maintaining relevance requires addressing the risk of scale mismatch, whereby solutions designed at one level, whether organisational, national, or planetary, can create new vulnerabilities at another. An organisation that builds internal antifragility but operates within a brittle international system remains exposed. A country that invests in pandemic preparedness but depends on global supply chains for the inputs that preparedness requires remains agentially constrained.

The practical implication is that resilience-building cannot be siloed by sector or scale. Public and private actors face the same overarching risks, and the session surfaced the value of bringing diverse actors into conversation to address challenges collaboratively. Here, input from the private sector regarding operational constraints and incentive structures provides a practical grounding to discussions that might otherwise remain abstract, ensuring that collaborative solutions remain relevant in a time of global change.

Opportunities

Despite pervasive anxiety, the seminar identified the opportunities that can emerge if the right conditions are created. At the organisational level, this begins by recognising that many established companies with deep domain expertise and large user bases have developed their own AI offerings, but these specialised models have not kept pace with frontier systems in general reasoning ability. Incumbents who fail to leverage their unique data, platforms, and domain knowledge effectively risk being outpaced by general-purpose AI tools that lack their specificity but exceed their capability. Equally, this represents a significant opportunity for companies that can integrate frontier AI with proprietary domain strengths, and for individuals who can position themselves at that intersection. This intersection points toward a deeper, more embedded form of AI adoption ahead, as advanced adopters move beyond using off-the-shelf tools toward training and customising their own agents for domain-specific tasks. The discussion of agentic AI surfaced interest in carefully scoped applications, such as investment management tools and workflow automation, where the boundaries of autonomy are well-defined and human review is built in. Further, AI has the power to answer highly specific questions for niche communities and contexts, with this democratisation of knowledge holding real potential if accuracy can be assured.

To maximise the effectiveness of this organisational adoption, leaders can draw on two concrete heuristics adapted from Hans Rosling's *Factfulness*. The first is to treat every AI statistic as a ‘lonely number’: ask what it was a year and ten years ago, and what it is per person, per company, or per dollar of revenue. The figures that actually drive good decisions tend to be rates, such as time saved per ticket or error rate per thousand documents, not headline totals. The second is to resist the ‘sleepy pilot’ explanation: ‘the data was biased’ is rarely the end of the inquiry, and the more useful questions are upstream; who chose this data, what was the model optimised for, and who does it serve? The work that makes AI adoption effective is usually done not by a heroic founder or model but by the unsung institutional and infrastructural people inside an organisation.

A second opportunity is collective verification. The wisdom of crowds is not really about crowds being wise; it is about different people being wrong in different directions and cancelling out, and groupthink is what happens when that independence collapses. Real-world applications demonstrate that conflict mapping networks only function when reports come in independently, and even well-meaning coordinated reporting broke it. Andreas's work on Community Notes shows the same principle at platform scale — its bridging algorithm only surfaces a note endorsed across the political spectrum, and such notes have proved comparable in accuracy to professional fact-checkers, working best in collaboration with them rather than as a replacement. That collaboration between commercial interests and research-led, public-

facing work is exactly the kind of intersection the wider Bridgehouse programme and its delegates can drive.

Beyond verification, there is also a clear opportunity to stop treating hallucination as purely a defect. Evaluated through the lens of the history of art, there are several cases where hallucination becomes productive: a scientific hypothesis is essentially an evidence-based hallucination; asking a model to write about a US president and getting a male default reveals that useful invention sometimes means going against the training data, not with it; and the diffusion models behind tools like Midjourney work by a kind of reverse-pointillism, recovering form from noise. Plausibility, not only literal factuality, is where much of AI's creative value lies.

This creative capacity highlights emerging possibilities for more community-centric and participatory approaches to AI, framing innovation as an opportunity to move beyond exclusive, extractive models of development toward futures where technological systems are co-created with, rather than imposed upon, diverse populations around the world. In healthcare, a Gates Cambridge Scholar discussed how AI-powered risk prediction tools may support earlier identification of high-risk individuals, more efficient allocation of healthcare resources, and more targeted preventive interventions. Particular attention was given to the potential impact in LMIC settings, where scalable and lower-cost prevention strategies could improve access to care and reduce unnecessary investigations for lower-risk populations. The seminar also explored how AI-supported communication tools may improve health literacy and patient engagement by explaining risks and preventive actions in simple, locally relevant language.

Other examples of this participatory shift include long-term ethnographic collaborations with artists, climate practitioners, refugee communities, and national AI initiatives developed in partnership with organisations such as the United Nations Development Programme and Microsoft. These initiatives demonstrated how AI can become a tool for cultural expression, collective agency, and social imagination when communities are actively involved in shaping technological systems. Realising these opportunities, however, depends on governance keeping pace with capability. Delegates raised the absence of a global regulatory standard, expressing concern about the lack of international collaboration on safe AI development and the risk of a fragmented landscape as ‘AGI’ approaches. The group also noted structural inequalities in who benefits: AI infrastructure demands are intensifying energy pressures, job displacement is falling disproportionately on certain groups, including women, and access to capable AI tools remains highly uneven. Inclusion by design, i.e. ensuring that AI systems serve broad populations rather than reinforcing existing hierarchies and epistemic injustices, was identified as both an ethical imperative and a practical condition for AI to deliver on its promise.

Finally, the discussion on polar climate engineering illustrated both the scale of interventions being considered and the ethical complexities of adopting them. The range of proposed techniques, from reflective aerosols to re-animating mammoths, represents a significant expansion in the toolkit available to those trying to slow climate breakdown. However, the session highlighted that an expanding toolkit does not necessarily simplify decisions. Interventions at planetary scale raise questions about governance, consent, and unintended consequences that existing international frameworks are under-equipped to handle. Many of the most advanced intervention technologies are being developed by private actors operating in a regulatory environment that proceeds far slower than these growing capabilities, demonstrating the disjunct between the public and private spheres.

The practical takeaway is that opportunity and risk are inseparable when considering such consequential decisions. Organisations and policymakers engaging with climate intervention need to build the governance capacity to evaluate proposed interventions critically, engage affected communities meaningfully, and resist framing any single technology as a solution that overshadows the need for wider systemic change.