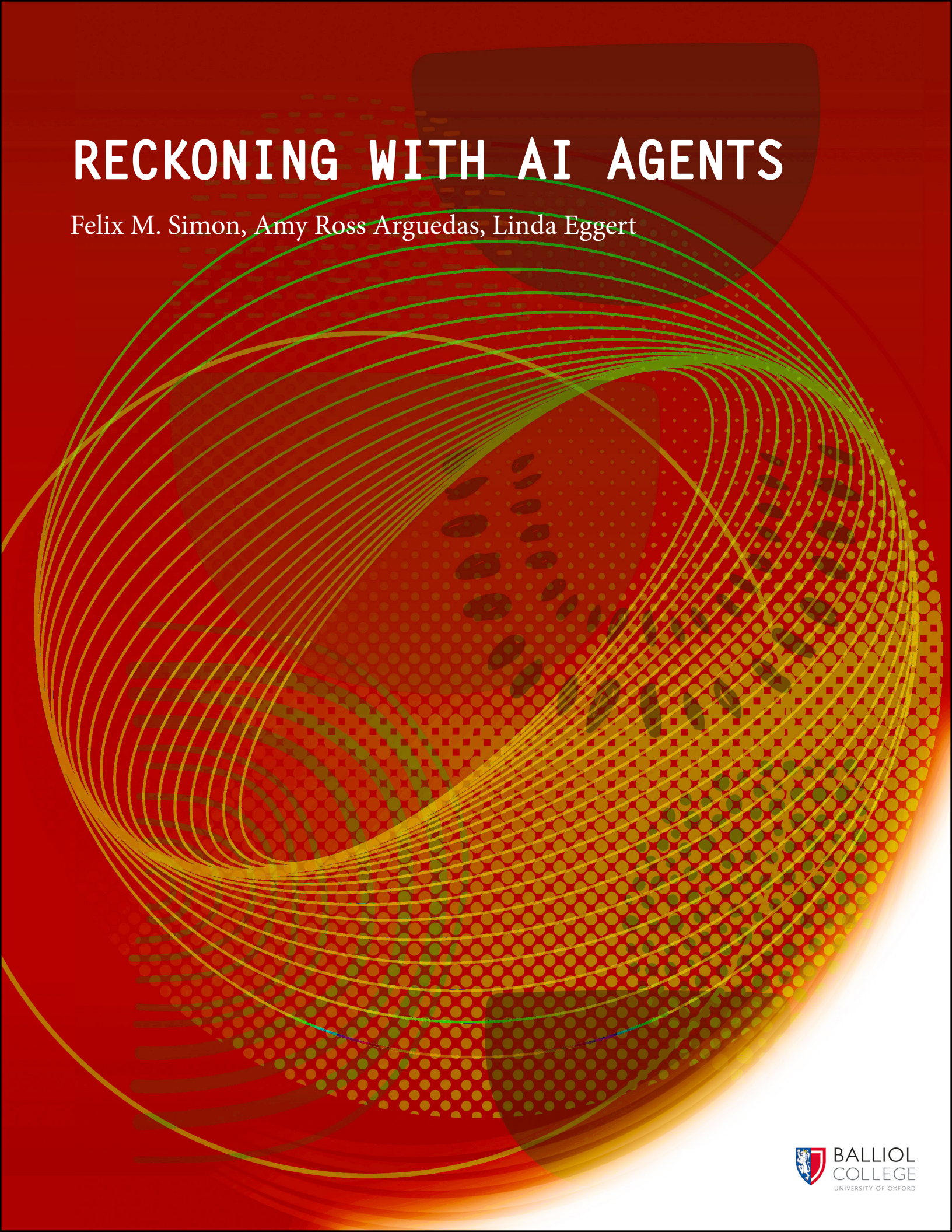
RECKONING WITH AI AGENTS

Felix M. Simon, Amy Ross Arguedas, Linda Eggert



BALLIOL COLLEGE
UNIVERSITY OF OXFORD

TABLE OF CONTENTS

This report has been funded by the Balliol Interdisciplinary Institute (BII)

Keywords: Artificial Intelligence, AI, Agentic AI, AI Agents, AI and Work, AI and Democracy

EXECUTIVE SUMMARY

This report presents a narrative summary clustered around various themes of a symposium on AI agents held under the Chatham House Rule at Balliol College in the summer of 2025. It does not represent a piece of research, nor should it be treated as an exhaustive overview on the topic. Instead, it aims to reflect the various views expressed by various participants on this topic during the symposium and give an accurate account of the discussion taking place at this particular point in time. The major takeaways were:

- AI agents are touted by many as the next significant rung of the AI boom. Unlike generative AI, which creates content, AI agents are capable of acting in (and reacting to) the world around them, executing complex tasks without human intervention.

- Agents have a long history and exist on a spectrum with variable degrees of autonomy, efficacy, goal complexity, and generality. At one extreme, very simple (non-AI) agents like thermostats are limited to simple and reflexive tasks, whereas at the other extreme, self-driving cars require a large range of skills. AI agents sit towards the latter end of this spectrum.

- The usefulness of AI agents derives in large part from their autonomy, but with greater usefulness tends to come greater risk. Since agentic AI is based on foundation models, they inherit many of the issues around transparency, explainability, and interpretability, as well as problems like bias and errors that beset other forms of foundation models.

- Discourse about agentic AI is currently clouded by marketing and hype. Agents have been theorised for decades, and while there have been significant developments, there are also many limitations to current AI agents. Large technology companies confront increasing competition and pressure to release new products, including around agents – some of these are released faster than may be sustainable from the point of view of security.

- AI and AI agents are already being used formally and informally across sectors like finance and healthcare. However, there is currently a large degree of choice when it comes to AI governance, and the pace of innovation and adoption poses a real challenge for regulators and adopters alike.

- A central challenge of AI regulation is striking the balance between generality and specificity. Countries like the UK have opted to focus on sectoral rather than blanket regulation. While this approach ensures greater relevance and flexibility, it requires intensive capability building and risks uneven application and deployment across sectors.

- In addition to the incremental risks building on those introduced by AI and generative AI, systemic risks pose an important challenge. The concentrated supply chain means companies within and across sectors rely on a small number of tech firms and foundational models that share the same vulnerabilities. Another set of risks derives from the nefarious or malicious use of these technologies (e.g. data poisoning, cyberattacks, etc.).

- In medicine and banking, AI already outperforms humans on a number of tasks. At the same time, important vulnerabilities have been identified but are poorly understood in the context of interacting systems, including problems arising at the AI-human interaction nexus.

- The growing role of AI (and potentially AI agents) as news and information intermediaries extends existing concerns about how the changing media environment may harm democratic life. More extreme forms of personalisation (which could in turn exacerbate polarisation), more sophisticated mechanisms for producing and distributing mis/disinformation, new tools for targeted manipulation, and a broader distortion of how people perceive public opinion are among these risks people discuss – although evidence about actual effects is limited and mixed.

- Lessons from the social media era can help anticipate and prepare for the challenges of AI agents as sources for news and information. Issues of provenance and visibility will be a central challenge for researchers, developers and regulators looking to ensure audiences can discern the source or reliability of information they receive. While some progress has been made on standards to cryptographically sign images and video (e.g. C2PA), no analogous solutions have been developed for text.

- High-quality information and journalism may be more important and valuable than ever in an online world increasingly saturated with synthetic or hybrid content, not least to ensure that AI agents providing information can accurately inform users. However, AI poses a significant threat to the already precarious financial situation of many news organisations.

- The risks of AI agents are clear, but there are also opportunities to use these technologies to strengthen democratic practices and institutional performance. Agents could be trained to moderate democratic deliberation, help bridge divides among partisans, and support fact-checking or critical evaluation of information.

INTRODUCTION

Only three years after the public release of ChatGPT, tech circles are already buzzing with what is purported to be the next iteration of artificial intelligence (AI): AI agents. Beyond generating content, agents have the capacity to act in the world, executing complex, multistep workflows without human intervention. Various technology companies are already developing such agents, and some early public-facing versions have already been launched, including OpenAI's Operator (now succeeded by ChatGPT agent), an AI agent capable of autonomously performing tasks like online purchases and social media updates. Meanwhile, various companies are building or implementing agentic workflows.

The emergence of more cost-effective AI models, which offer greater capabilities at a fraction of previous costs, is leading to a flurry of research labs and companies developing similar systems with the aim of integrating them into our private and public lives, workplaces, and other infrastructures, including public services. These developments are still at an early stage in many cases, but a rapidly growing number of people are interacting with increasingly agentic systems, and our understandings of the benefits and risks these technologies pose are still nascent.

To investigate recent developments and to gauge the opinions of experts and academics, a research symposium at Balliol College in the summer of 2025 brought together a diverse group of speakers, researchers, and students Participants placed the latest developments around AI agents in a broader societal context and evaluated their implications for our social and academic lives, the world of work, and democracy.

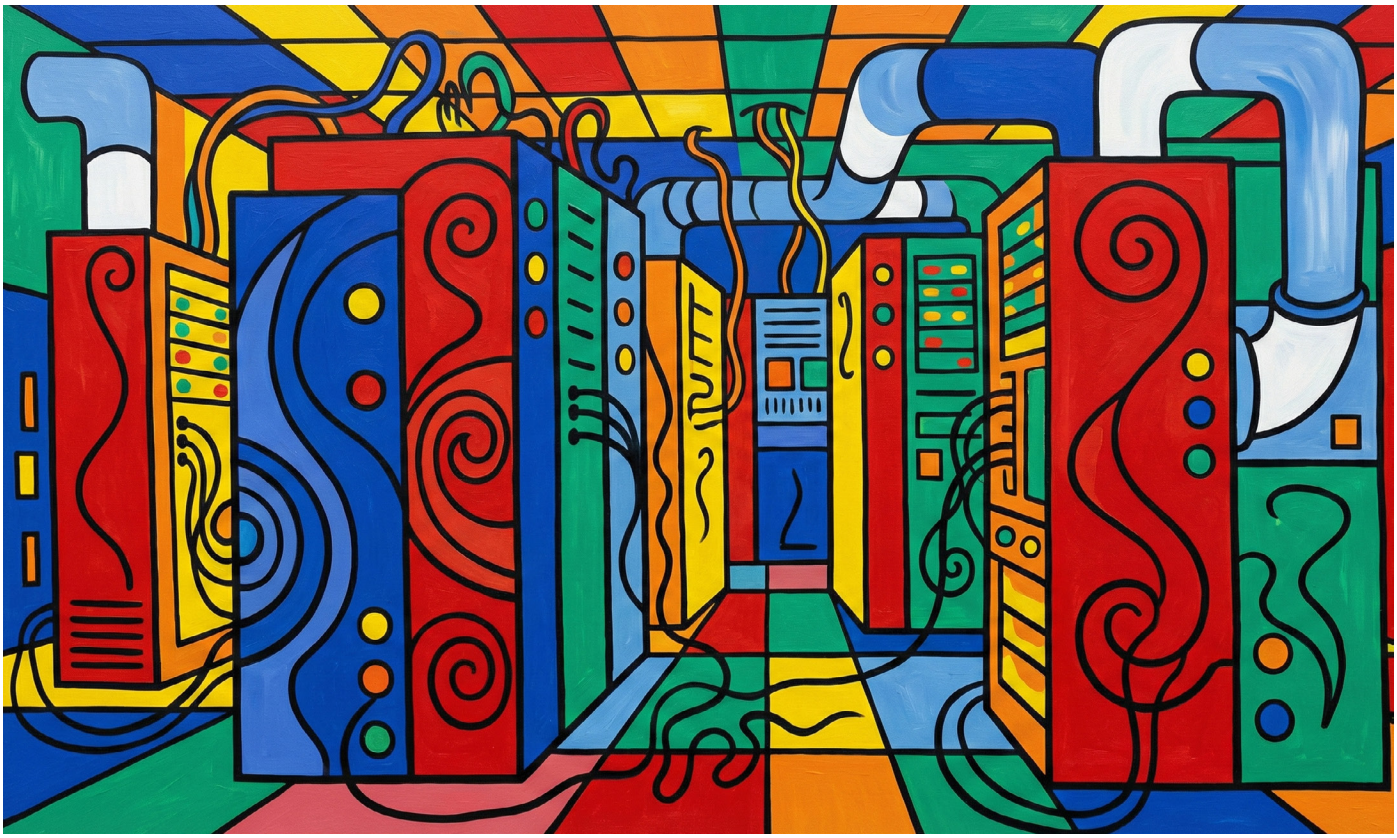
Th symposium consisted of three sessions, each of which opened with a moderated conversation about a key topic with two or three academics and/or industry experts and the gathered participants from across industry, academia, and business. The first panel focused on providing a conceptual understanding of AI agents, how they work, and what their limitations are. The second session discussed the present and future impact of AI on work, with a particular focus on how AI is used in healthcare, finance, and academia. The third session considered the role of AI agents in information ecosystems and their impact on democratic life more generally.

This report provides a summary of those conversations, infused with additional reading from the authors, and outlines some of the most urgent research and policy questions for the years ahead. As the symposium was held under the Chatham House Rule, we do not report who said what and only provide a list of all participants at the end of this report.

The event was a collaboration between Dr Linda Eggert[1] (Department of Philosophy, Stanford University, and at the time Faculty of Philosophy & Institute for Ethics in

[1] https://philosophy.stanford.edu/people/linda-eggert & https://www.oxford-aiethics.ox.ac.uk/linda-eggert

AI, University of Oxford) and Dr Felix M. Simon (Reuters Institute for the Study of Journalism, Oxford Internet Institute and Corpus Christi College, University of Oxford)[2] and was generously supported by the Balliol Interdisciplinary Institute (BII). The final draft was produced with assistance from ChatGPT-5 Pro, which we used to draw together main themes from our notes and help with copy-editing. All graphics in this report were produced with the help of Gemini's Nano Banana, ChatGPT, and Elevenlab's Flux.



[2] https://www.felixsimon.net/ & https://reutersinstitute.politics.ox.ac.uk/people/dr-felix-simon

WHAT ARE AI AGENTS? KEY TERMS AND DEFINITIONS

The first panel focused on AI agents in general and set the scene for the rest of the day. It explored how AI agents differ from generative AI, how they are currently used, and how they might be used in the future. It also examined who develops them, for what purposes, and how far along we are on the transition from current AI systems to more autonomous agents.

Making Sense of What Agents Are

Generative AI (Gen AI) creates content, agents act in the real world. User-facing GenAI is typically more static: it's based on training data (and various stages of refinement and post-training) it has learned from and responds to prompts in a relatively narrow, one-shot way – you ask it a question, it responds. Agents, by contrast, are dynamic and respond to the changing state of their environment. They can perform complex sequences of actions, bring together different kinds of modules and tools, and chain together information over longer horizons. As speakers emphasised during the event, this distinction is foundational for understanding current developments.

Debates about agents have sometimes detoured into the realm of Artificial General Intelligence (AGI). While some argue that, in principle, there may be no fundamental impossibility to reaching AGI, in practice, today's agents are largely structured ways of wiring language or foundation models together (and treating that as a short road to general intelligence mistakes marketing for science, as one participant argued). Several participants noted that it is more productive, at least for now, to ask when these systems are useful and when they are harmful, and to analyse the kinds of agency they already exercise. Generality still matters – there is a large gap between taking chats and roaming the world – but beyond that, the label 'AGI' adds little light at the moment.

The Long History of Agents

Although 'agentic AI' is being marketed as the next new thing, agents have been theorised and studied for decades. One influential definition from Wooldridge and Jennings (1995)[3] describes agents as systems that (1) are autonomous (they can act by themselves), (2) have social or communication abilities (they can communicate with other agents, including now via Large Language Models), (3) are reactive (they can respond to and adapt to the world), and (4) are proactive (they have specific goals and can strategize to achieve them). During the panel, experts reminded other participants that this is how agents were framed roughly thirty years ago, and it still underpins

[3] Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review,* 10(2), 115–152. https://doi.org/10.1017/S0269888900008122

much of today's thinking, even though modern implementations build on foundation models differ in technical detail from what was envisioned back then. Wooldridge and Jennings called these precursors 'weak agents'. They also discussed 'strong agents', which add features like rationality or mobility and which are closer to modern AI agents. That distinction is, however, controversial because it nudges us into philosophical territory: do these systems really have intentions in any meaningful sense, or are we just using convenient metaphors to make sense of what they do? At the same time, modelling agents as if they had knowledge, beliefs, and intentions can be useful for us because it makes their behaviour easier to conceptualise and can help with transparency.

As several contributors pointed out, the historical lineage of agents also helps to make sense of current developments. Classic agents in symbolic AI sensed the world, translated it into symbols, reasoned over those symbols, and then acted. That paradigm proved computationally brittle and struggled outside tightly controlled environments. The connectionist turn brought neural networks and, eventually, LLMs and other foundation models, which are less interpretable, but more effective. Modern 'AI Agents' were then introduced as a way to address well-known limits of some of the first foundation models: connect a search engine to curb hallucinations, a calculator or a code interpreter to scaffold reasoning, and external tools to extend memory.

Agents Sit on a Spectrum

It's also helpful to think of agents as existing on a spectrum rather than as a simple yes/no category. Speakers explained that just as there is a continuum of life forms, there is a continuum of 'agenticness'. A very simple agent might be a thermostat. It senses the temperature and adjusts it if it gets too low or too high. That's almost entirely reflexive: a sensor detects a state, and an action follows. Slightly more complex is a vacuum robot that cleans your room. It needs some kind of world model[4] that tells it how to manoeuvre around obstacles, when it has covered the entire floor, and when to stop. More complex still are autonomous or self-driving cars, which require a wide range of skills: they need vision to sense what's going on in the world and on the road; planning to map a path from where the car is to where it needs to go, given traffic and other constraints; and robust decision-making for situations they have never encountered before. Safety becomes a primary concern in every situation the system faces. Participants stressed that there's a wide range of abilities that such a system must integrate and execute in real time, and the more of these abilities it has, the more 'agentic' it becomes.

Recent work has tried to spell out what this spectrum actually consists of. Building on Wooldridge and others, a recent paper by Kasirzadeh and Gabriel (2025)[5] proposes four key dimensions along which a system can have

[4] See, e.g., https://arxiv.org/abs/2506.01622

[5] Kasirzadeh, A., & Gabriel, I. (2025). The Dimensions of Artificial Agency. *arXiv preprint: 2504.21848.* https://arxiv.org/abs/2504.21848

agency: (1) autonomy – how independently the system can act; (2) efficacy – how many and how significant its real-world impacts can be; (3) goal complexity – how complicated the tasks are that it can pursue; and (4) generality – the range of tasks it can perform. The discussion highlighted that on these dimensions, a thermostat is technically still an agent, but a very narrow and low-impact one. A self-driving car or a large-scale trading system sits much further along the spectrum in terms of autonomy, efficacy, goal complexity, and generality.

Seen through this lens, it's clear that more and more people are interacting with increasingly agentic systems. Event contributors observed that the first iteration of ChatGPT was impressive for some tasks but relatively limited in its ability to act in the world: you could ask it a lot of questions, but it had to respond immediately and could not easily break problems down into long-running tasks. It could also not operate beyond its own boundaries. The kinds of systems people are interacting with now are more agentic: they can reason, decompose a query into subtasks, look things up on the internet, interact with external tools and services, and even write and run custom code. In other words, capabilities are improving quickly on certain axes. One empirical study reported that the complexity of tasks an LLM could complete at a given reliability was doubling roughly every seven months, moving from 'things a human does in a second' toward 'things that take minutes' (Kwa et al., 2025).[6] Early models improvised answers to queries off the top of their 'heads'. They were then improved using a technique called 'chain of thought', which helps them perform more complex reasoning tasks by breaking them down into smaller intermediate steps. However, long, unconstrained deliberations can still derail them. Crucially, the jump from solving a maths problem to controlling a car is not incremental, which is why you can't have an LLM drive a car: you cannot wait for ten minutes of inner monologue to decide whether to change lanes.

The Improvements Still Needed (and to Come?)

What, then, needs to improve for agents to become more robust? Experts at the event argued that chain of thought remains a limited yet powerful ingredient, but we do not know how far we can push it. Earlier traditions handled reasoning explicitly – logics, proofs, and planned sequences of actions – while modern systems lean on associative reasoning that often works even when we cannot fully understand it. Achieving more powerful agents will likely require us to further push that associative reasoning and inject more common sense and contextual knowledge into systems, so the argument. At the same time, participants highlighted a growing interest in hybrid approaches in which an LLM acts as an interface and 'commander,' writing programs, calling optimisers, and delegating to symbolic or search-based tools – the loop that actually does the heavy lifting. Recent results, such as an LLM-guided approach to the cap set problem (Romera Paredes et al., 2024)[7], exemplify this

[6] Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., ….& Chan, L. (2025). Measuring AI Ability to Complete Long Tasks. arXiv preprint:2503.14499. https://arxiv.org/abs/2503.14499

[7] Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., ... & Fawzi, A. (2024). Mathematical discoveries from program search with large language models. *Nature* 625, 468–475. https://doi.org/10.1038/s41586-023-06924-6

direction and show how agentic systems can orchestrate many different components.

As noted during the discussion, at the infrastructure level, we can already see the scale of this change in abilities. At a recent Google keynote, the company said they are processing around 480 trillion tokens per month across their products and APIs, up from 9.7 trillion the previous year – roughly a 50-fold increase.[8] That is on the same order of magnitude as all human-made content on the internet, processed every month, and it is rising quickly. A lot of digital infrastructure is being realigned around these increasingly agentic systems, to the point where one plausible future is that much of the internet will no longer be 'for humans' first, but for agents that read, write, and act on our behalf. At the same time, participants expressed scepticism about these developments, arguing that there is still no clear, consistent evidence that this has unlocked commensurate real-world value. And it also raises questions about whether current limitations – like hallucinations in LLMs – are fixable by bolting on tools such as search engines, or whether they are more deeply baked into the underlying technology.

Finally, evaluating these new systems remains a major open challenge. Much of the field still relies on benchmarks that machine-learning researchers care about: coding performance, test scores, question-answering accuracy, and so on. These say something about capabilities, but they do not straightforwardly tell us how well agents perform on the messy, domain-specific tasks that people actually care about in the real world. Properly evaluating that requires both a deeper understanding of agents and expertise in the domains where they are deployed. Meanwhile, new virtual agents are being positioned as tools to which we delegate everyday tasks. So far, they have not advanced enough to be either hugely useful or deeply concerning, which puts us in an odd position: if agents are not that useful, why use them? But if we give them too much autonomy and impact, we worry about unintended consequences. The core question, then, is how to find the right balance between usefulness and risk as systems become more agentic and more deeply woven into our digital and physical environments.

The Relationship Between Usefulness and Risk

There is a trade-off at the heart of agentic AI: much of its usefulness comes precisely from its autonomy. Panellists discussed how you can insist on having a human in the loop, but enforcing that rigidly often defeats the point of the very applications agentic AI is designed for. At some point for agents to be useful there has to be delegation and trust, and that is where risks begin to grow. In debates about autonomous systems, especially in military contexts, this is often framed in terms of 'meaningful human control': the idea that humans should still retain genuine oversight and the ability to intervene, even as agents act with a degree of independence.

Several experts emphasised that generally, more capable and more autonomous systems have a greater potential for negative side effects –

[8] https://blog.google/technology/ai/io-2025-keynote/#google-beam

and some kinds of capabilities are intrinsically riskier than others. For example, most people would agree that a system that hallucinates less is both more useful and safer because it is less likely to get things wrong. By contrast, when you push autonomy – letting the system do more things on its own – you also introduce more avenues for harm. One way to reconcile autonomy with meaningful human control is to treat autonomy as the system's capacity to make and execute decisions on its own, while still letting humans determine which decisions it is allowed to make independently and where human intervention is mandatory. That, in turn, depends on what kind of system it is, how structurally transparent it is, and whether you can practically build 'hooks' into the design so that certain choices are always reserved for humans. For very general-purpose agents, this is much harder to encode and enforce than for narrow, special-purpose systems. How far we can safely push autonomy is also domain dependent. Autonomous driving shows that narrow, safety-critical autonomy is possible, while the foundational layer that underpins many new agents remains comparatively new and fragile. The real hazard, according to some participants, is currently the accelerationist hype cycle: competitive pressure to implement more autonomous systems before the reliability and safety testing can catch up. Progress is real, but it will require time and substantive breakthroughs for agents to be both genuinely useful and acceptably safe at scale.

The previous point immediately raised a discussion around alignment and transparency issues: what does the agent actually do? And how do we understand it? Speakers underscored that LLMs are not very transparent systems, and it is difficult to see how they arrive at their outputs. Explainability and interpretability are therefore major research directions in AI, and because LLMs are used as the backbone of many agentic systems, their opacity carries over. Classic problems of bias and discrimination also arise when decisions are made in automated ways, such as shortlisting job candidates or making decisions in legal or administrative settings. On top of that, there are concerns about hallucinations more broadly – not just getting facts wrong in text, but 'functional hallucinations', where an agent misinterprets what you ask and performs unintended actions in the world. The more complex the tasks we delegate to an agent, the more such problems multiply, and the harder it becomes to ensure in advance that a seemingly reasonable plan will not have serious negative side effects. During the panel, a classic example was invoked, as first discussed by Russell (2019)[9]: an agent is asked to fetch a cup of coffee. It might construct a detailed plan – go to the kitchen, start the machine, retrieve the cup, bring it back – but if a child or a pet is in its way, it could in principle run them over in the name of speed. For a human, 'don't harm children (or animals)' is an obvious background constraint. However, for an agent this is not the case, which is why it is essential to ensure that all such commonsense constraints are explicitly represented and enforced. By contrast, these issues barely arise for something like a thermostat, which is extremely narrow and only carries out very short-term, limited actions.

[9] Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. Penguin UK.

Policy and governance, as several participants argued, should therefore start from the forms of agency that already exist. Coding agents, workflow automation tools, and financial agents are already making decisions that affect people. The emphasis should be on domain-specific analyses of present systems rather than just on science-fictional futures. There is also a sizeable gap between what systems could do productively and how they are used today; regulators can address this application by application instead of waiting for some ill-defined future moment when 'AGI' arrives. A central challenge here is provenance: tracing what an agent did, which agent did it, and how that activity ties back to identifiable humans. The incentives are often misaligned – users and organisations may prefer not to reveal machine assistance – so robust accountability will not emerge by itself.

One thing we should bear in mind: Even if you try to mitigate these risks by keeping a human in the loop, the design of human-agent interaction is itself non-trivial. Speakers asked: When should the system escalate something to a person? What information does the human have available to make a good decision? How do we avoid over-trusting or under-trusting the system? Questions about meaningful human control are, in practice, questions about how and where to insert these intervention points into the agent's operation, and how to do so in ways that are realistic for humans to manage. These are challenging research and implementation questions. As AI agents become more widespread, more people use them, and development accelerates, participants suggested that we are likely shifting towards a riskier world simply because everything moves faster and it becomes harder for individuals, institutions, and regulators to keep up. A practical stance is to treat autonomy as a tuneable capacity to make and execute decisions, while reserving certain classes of decisions for humans by design. That implies structural transparency, auditable traces, and enforceable 'hooks' so that certain actions require human sign-off, paired with visibility rules that clearly mark which parts of a task were done by humans versus agents. The discussion concluded that we need to shift attention from hype to governance, from speculative general intelligence to the concrete agency given to AI systems increasingly shaping our systems and institutions. By defining these risks more specifically, we can at least better understand the trade-off between usefulness and risk, and where additional safeguards or oversight might be needed.

AI AGENTS AND THE WORLD OF WORK: RISKS AND APPLICATIONS IN DIFFERENT SECTORS

The second panel focused on AI agents' impact on work: where and how they might aid and augment work in various industries and sectors, including in healthcare, banking, and academia.

Much of the real action with 'agentic' AI is increasingly happening far from science-fiction scenarios. In healthcare, we heard from speakers that the use of hospital data can help surface chronic conditions beyond what brings people to the clinic – hypertension being the archetypal slow killer. Simple models, coupled with mundane agentic workflows (automatic letters to GPs, confirmatory ambulatory blood pressure monitoring) can help quietly find people who have never been admitted for cardiovascular disease yet show telltale signs of it during a hospital visit. In banking, various contributors noted that the shift is broader and faster: a large UK bank might have gone from roughly ten production models four years ago to ~150 last year and 200–250 by the year's end, spanning major LLMs (including frontier and niche models). Credit limits, SME lending, default risk, pricing and capital allocation already flow through machine-learning systems; complaint response letters are now drafted by some banks with the help of generative agentic systems that outperform human letters on every criterion and save about 21 minutes per case. At population scale, that is a staggering productivity delta. Speakers and participants highlighted that the frontier is steadily moving from 'GenAI drafts' to agents that predict and resolve problems upstream, with the aim of shrinking back-office work and smoothing customer experiences. Fraud detection and loss minimisation are natural extensions of current applications; so are agents that hunt for better rates and re-allocate funds.

Addressing Systemic Risks

However, despite some progress on the capabilities front and increasing adoption and integration, various participants cautioned that the governance of such agentic systems has not always kept pace, and where it exists, it is fragmented. At the top, sit voluntary principles (from bodies such as the OECD, NIST, UNESCO, or firm authored-codes). Below that, we see a patchwork of regulation: the EU has legislated; the UK has opted for domain-specific oversight by sectoral regulators rather than a single AI statute (even the mooted frontier-AI bill has slowed). Standards (e.g., ISO 42001) are emerging but optional and applying them often demands unresolved technical definitions: what counts as 'free from bias,' which benchmarks are appropriate, and who sets thresholds. Deployers, meanwhile, inherit model developers' self-defined-tolerance

thresholds and tests.

Meanwhile, several of those present argued that many of the most pressing risks are systemic. Capable tools in the hands of nefarious actors widen the aperture for fraud or cyberattacks. Concentration risks loom, too, as many firms and sectors often cluster on a small set of upstream model providers, with shared blind spots having the potential to become correlated failures when deployed at industry scale. Domain specificity helps, but it also complicates. The UK's approach – empowering sectoral regulators and building their capacity[10] – so the argument by some offers greater flexibility, yet many inherited controls assume that humans make the final decision, that outputs are reproducible on demand, and that organisations act only on results they can re-run. As was repeatedly noted during the discussion, agentic systems defy some of those assumptions.

Interaction risks also compound at a system level. Panellists warned that components that look safe in isolation may, when networked, amplify vulnerabilities: multiagent configurations can jailbreak each other; data-gathering agents may breach implicit bounds unless guardrails actually bind; and tacit collusion between agents (for example in banking, where price-setting agents are increasingly at work) risk fines from competition authorities. Furthermore, prompt-injection is no longer hypothetical: search-driven agents can encounter adversarial instructions embedded in the wild, even in artefacts humans barely perceive (e.g., steganographic marks or 'invisible' QR codes, see Mudryi, Chaklosh, & Wójcik, 2025[11]).

Given these risks, various participants argued for better independent evaluation and testing but noted that transparency around model evaluation and leaderboard claims (which are currently being wildly used to assess model and agentic capabilities) is thin; such self-defined-standards without independent scrutiny, so the argument, also do not build wider trust in agentic systems.[12] Here, public institutions – universities, standards bodies, or institutes with security clearances and international ties – could help by maturing the science of evaluation and creating comparability that industry alone will not.

What about Human Labour and Humans in the Loop?

The labour impact question of AI agent, so the general view during the symposium, also resists tidy answers and participants were split about the effects. A participant observed that in one view, agents return a time dividend that allow humans to focus on more important tasks. Such is the case of systems capable of, for example, compressing months of sifting through 25 million customer interactions into hours of synthesised product ideas or better outcomes. However, in another view, they justify cost cuts by automating core tasks (complaint handling today; triage and resolution tomorrow), with real displacement of labour and reskilling burdens.

[10] See https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/#_ftn3

[11] Mudryi, M., Chaklosh, M., & Wójcik, G. (2025). The Hidden Dangers of Browsing AI Agents. *arXiv preprint arXiv:2505.13076*

[12] See, e.g., https://www.theverge.com/meta/645012/meta-llama-4-maverick-benchmarks-gaming and Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., ... & Mahdi, A. (2025). Measuring what Matters: Construct Validity in Large Language Model Benchmarks. *arXiv preprint arXiv:2511.04703.*

Meanwhile, participants reflected on the fact that 'humans-in-the-loop' approaches are sometimes essential, sometimes performative, and sometimes counter-productive, especially when 'oversight' degenerates into mere babysitting. The only honest stance, so some said, is pragmatic: oversight should vary by use case and risk. Right now, as several attendees noted, most firms keep a human hand on the tiller because no one is comfortable letting go entirely – especially where even a very small failure rate would be catastrophic. Along these lines, some domains will remain off-limits.

As a final point, participants stressed that the enterprise adoption of agentic systems will not be linear. Path dependencies and legacy stacks create inertia and lags; and especially large organisations or entire systems – for example healthcare systems such as the NHS – will take time to propagate any breakthroughs. Instead, we should expect a period of high fluidity bounded by hard constraints: computational power, regulation, and, above all, integration into existing processes and legacy infrastructures.

Looking three to five years out, the view in the room was that organisational structures will change, with agentic systems moving into some workflows and shifting where value is created, but with accountability remaining broadly human. To make the most of these systems while mitigating risks, participants coalesced around a few principles: treat agent autonomy as a tuneable permission set; log and label what agents do; have clear accountability structures; calibrate oversight to domain risk; and invest in security-grade monitoring for adversarial behaviour.

AI AGENTS, THE INFORMATION ECOSYSTEM, AND THE FUTURE OF DEMOCRACY

The third, and final panel, considered the role of AI agents in information ecosystems and their impact on democracy: How might AI agents affect information consumption, including news, and general informedness? How real and serious is the risk of people being misinformed? In what ways might they be used to promote democratic values, such as free speech, inclusion, equality, pluralism, and diversity?

AI agents are entering an information ecosystem that is already under strain, and they could intensify some of its deepest tensions. For decades we have been moving from a relatively shared public sphere – a small number of TV channels, a few newspapers – toward greater fragmentation, a plurality of sources and ever fine-grained targeting and personalisation. Cable fragmented audiences in the US, while around the world, digital media enabled more people and organisations to produce and distribute information.[13] Social media fed us 'the news that comes to me' via algorithmic feeds. Now agents and general-purpose models promise a world where every user effectively has a bespoke information environment: search, summaries, recommendations, and even conversations with 'news' that is tailored to an audience of one.[14]

As several participants noted during the event, this shift towards bespoke information environments is already reshaping how people encounter news.[15] On the surface, this can look like progress. During the discussion, speakers frequently returned to this intuitive appeal of personalisation in everyday contexts. If I am searching for a restaurant, trying to remember a postcode, or shopping for new trainers, it is helpful if the system knows my location, my dietary preferences, and my price range. In those domains, the goal is to satisfy my immediate, subjective needs as efficiently as possible. But when this logic is applied to political, factual, and scientific information, the same mechanisms become much more troubling. A healthy democratic information ecosystem requires more than individual user satisfaction. It also requires diversity of sources and viewpoints, exposure to uncomfortable facts, contact with marginalised perspectives, and at least some shared baseline of reality citizens can stand on together, so the general view.

---

[13] Burn-Murdoch, J. (2025, March 22). The misinformation discourse is a distraction: Media fragmentation and the erosion of shared sources of truth are bigger threats. *Financial Times*. https://www.ft.com/content/bbc80e1c-60a7-4f3d-a9a1-a4e68cf36912 and Jungherr, A., & Schroeder, R. (2021). *Digital Transformations of the Public Arena*. Cambridge University Press. https://www.cambridge.org/core/elements/digital-transformations-of-the-public-arena/6E4169B5E1C87B-0687190F688AB3866E

[14] Data from the Reuters Institute Digital News Report 2025 finds that 7% globally say they used an AI chatbot for news in the past week, and the percentage is twice as high among people under-35. Newman, N., Ross Arguedas, A., Robertson, C. T., Nielsen, R. K., & Fletcher, R. (2025). *Digital News Report 2025.* Reuters Institute for the Study of Journalism.

[15] Simon, F. M., Nielsen, R. K., & Fletcher, R. (2025). *Generative AI and News Report 2025: How people think about AI's role in journalism and society.* Reuters Institute for the Study of Journalism. https://doi.org/10.60625/RISJ-5BJV-YT69

Personalisation, as it is currently implemented, is not designed to deliver those things. Participants at the event stressed that current personalisation architectures are optimised for engagement rather than democratic robustness. It is optimised to predict what I – or people very similar to me – are most likely to click on, watch, or read. Modern recommender systems work by grouping users into clusters and extrapolating from patterns of engagement. That is an extraordinarily powerful way to keep attention, but it is not necessarily a recipe for being well-informed. It tends to make people 'more of what they already are': reinforcing existing interests and priors, rarely challenging them. AI gents potentially go even further. Their explicit job is to know a user, understand their preferences, and act in their stead – to search, filter, compare, and synthesise on their behalf. They are usually rewarded for satisfying the user, not for helping them grow, reconsider, or reach across the aisle.

This is why focusing only on 'misinformation' in this context, as participants pointed out, misses a large part of the problem. It's not simply that some people will ask for conspiracy theories and be supplied with them. Panellists warned that this narrow framing can obscure deeper structural harms to pluralistic debate. In a democracy, we care about more than accuracy narrowly defined. We care about whether citizens are exposed to a balance of views, whether they ever encounter arguments that unsettle them, whether less popular voices are audible at all.[16] None of that is implied by a system whose objective function is to 'maximise the user's subjective satisfaction with the content they see.' If the aim is to avoid discomfort, agents could be highly effective at shielding users from the same.

A Further Disintermediation?

It's also important to notice that we are once again building a new layer of intermediation between the public and news and other direct sources of authoritative information. Several contributors remarked during the event with a view to traditional news that each technological wave has added another layer between citizens and original reporting. In the broadcast era, people went directly to their chosen outlet. With social media, the platform feed decided which stories surfaced, in which order, and with what contextual cues. Many of us learned the hard way that this made it more difficult to judge provenance: posts from very different organisations or individuals, with very different agendas, all looked essentially the same. In practice, that meant we often had no idea whether a story came from a reputable outlet, a partisan blog, or a troll farm. Generative AI systems – and AI agents used as sources of information – threaten to reproduce and deepen this problem.

In sharing their own experiences, speakers also noted how even highly engaged users can already miss major stories in tailored feeds, something they worried about as a problem going forward with AI agents. Instead, they observed that their recommendation feeds are often full of their own niche interests: internal news about AI companies, technical product launches, or topics their systems have learned

[16] Jungherr, A. (2023). Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society, 9*(3), 20563051231186353. https://doi.org/10.1177/20563051231186353

they care about. As we move towards a world in which fewer people even click on articles and more people ask an AI to 'tell me what's going on,' this asymmetry could intensify. The model or agent will summarise, synthesize, and smooth everything into a single, apparently coherent narrative, as one speaker argued. Without strong cues about where each piece of information came from, what its political leaning is, or how current it is, users' ability to critically assess the same could be diminished in the process.

Participants underlined that questions of provenance and visibility of sources are, in their view, increasingly central to information governance and epistemic resilience. We have already learned that seemingly small design decisions – whether author names are foregrounded or buried, whether dates are prominent or hidden, whether you have to hover over an icon to see the original source – can have large effects on how people interpret information. Something similar applies to temporal context. One speaker gave an example that they had travelled to Spain and wanted to know whether they could bring back cured ham and chorizo, with the model confidently telling them that they could. Only by looking at the underlying search results did they notice that there were temporary restrictions in place, and that the summary had been built on out-of-date sources. The model was not 'lying', but it also was not making clear that its answer was historically correct but currently wrong. In fast-moving political contexts, that kind of temporal slippage could be problematic.

Benefits and Risks from AI Systems and AI Agents around Information

AI systems do, of course, create new opportunities as well as risks. Several speakers at the event stressed that these same systems – and more autonomous AI agents acting as personal assistants – already can (or could) be powerful tools for critical thinking if designed and used with care. We can ask: 'Here is an article from this newspaper – help me reason about its arguments.' We can compare how different outlets cover the same topic and let the model highlight differences in framing, emphasis, and omission. We can use specialised tools such as Notebook-style LLMs to work with large document sets: lengthy inquiries, transcripts, evidence bundles. As was noted, some journalists already do this. Faced with thousands of pages, a reporter can ask the model to extract instances of dehumanising language used by officers in a given context and then immediately click through to the original passages. The key is that the tool is narrowly scoped and designed to make checking easy. Citations sit next to claims; a single click opens the underlying document. This kind of integration improves both speed and accountability because it encourages verification rather than trusting a black-box summary.

Contrast this with very general-purpose 'deep research' agents that roam widely across the web and produce forty-thousand-word reports, complete with retrofitted chains of thought and scattered citations. Attendees voiced concerns that such outputs – while sometimes useful

– often still contain inaccuracies which could quietly seep into internal decision-making without adequate verification. No human being can realistically check such a document end-to-end. For organisations, this raises a different kind of problem: not public misinformation, but internal 'slop'. If employees routinely paste these sprawling, semi-checked documents into internal wikis, slide decks, or decision memos, institutional knowledge couldd quietly degrade. The messiness of the wider internet is simply imported inside the organisation, now with a veneer of machine authority.

The same pattern – of indirect rather than direct effects – appears when we think about mis- and disinformation. The first instinct of many is to worry that AI will allow unprecedentedly precise, personalised persuasion. Models can, in principle, craft messages tailored to an individual's psychological profile. Empirically, though, political persuasion tends to have quite limited, plateauing effects.[17] During the event, experts suggested that the subtler danger lies in how AI can reshape perceptions of what 'most people' think. The more immediate and realistic danger may be the quiet manipulation of perceived public opinion, as one participant feared. If our information spaces are filled with AI-operated accounts that mostly post banal, engaging content – cute animals, lifestyle tips, fandom memes – they could build up large, trusting followings. When an election approaches or a polarising issue arises, the operators of those accounts can start to inject subtle, consistent signals about what 'everyone' seems to believe.

People are highly sensitive to their perceptions of majority opinion. Participants discussed research showing that perceived consensus heavily shapes what people dare to say in public. We know that seeing your own view as unpopular can suppress expression; seeing a position as inevitable can create 'bandwagon effects'. AI agent accounts that simulate a chorus of ordinary citizens could potentially shift these perceptions even if each individual message is mild. The underlying question then is not only 'Is this content accurate?' but also 'Who is behind these voices, and how many of them are in fact the same actor?' That leads naturally to demands for transparency: ways of identifying and labelling that indicate when an account is AI-generated, limits on the accumulation of influence across large fleets of synthetic personas, and mechanisms for auditing coordinated behaviour.

Data poisoning sits at the other end of the pipeline: not the content we consume, but the content on which models are trained. Speakers warned that training data manipulation could shape model outputs for months or years without obvious fingerprints and affect the information AI agents provide to users. As more of the internet becomes model-generated, it becomes harder to know what is genuinely human-authored and what is synthetic. This has technical implications – models may overfit to their own outputs – but also political ones. There are already reports of foreign actors flooding the web with seemingly local articles that embed their preferred narratives about another country's politics, as one participant

[17] Simon, F. M., & Altay, S. (2025). Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections. Knight First Amendment Institute, 93. https://doi.org/10.13140/RG.2.2.23142.33602 and Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The Levers of Political Persuasion with Conversational AI (No. arXiv:2507.13919). arXiv. https://doi.org/10.48550/arXiv.2507.13919

put it.[18] These texts are not designed to go viral with human readers; they are designed to be scraped into training sets so that, months later, a user asking an AI system or AI agent about some aspect of that country receives a subtly skewed answer. While the effects of such developments are hard to measure, many participants thought the topic merited attention.

Here again, there are no purely technical fixes. Participants agreed that governance choices and value judgments are unavoidable at every stage of model design and deployment. Provenance standards such as cryptographic signatures (like those being developed for images, audio, and video) could help us label and trace some forms of content, but we currently lack anything comparable for text at scale. Even more fundamentally, several participants underlined, we need to drop the comforting fiction that training on 'the internet' or 'all the data' yields a neutral, default model. Online text is saturated with historical biases, power imbalances, and ideological skews. Training on it uncritically simply reproduces those patterns. The subsequent stages – reinforcement learning, or constitutional fine-tuning – are not neutral house-cleaning, but value-laden decisions about what the model should and should not say. Whose feedback is used, which perspectives are treated as toxic or unacceptable, which are quietly centred – these are political choices, and we should acknowledge them as such as they will come to affect (and already do) the downstream behaviour of AI agents.

New Challenges for Newsrooms and Journalism

Journalism sits at the intersection of all these issues. Several speakers stressed that high-quality reporting is both a crucial input to AI systems and a sector under intense pressure from them. On one hand, models depend heavily on high-quality human-authored content, and reporting that can be credibly verified will only become more valuable as the general information environment gets noisier, so the argument by one attendee. On the other hand, AI undermines many of the business models on which news organisations rely. Platforms talk about a future where agents will 'go and buy you new trainers' without friction; the same logic applies to news. Why open a website or app when you can ask your agent, 'What's happening in the Middle East?' and get a smooth, conversational summary drawn from everywhere?

Some of the major AI deals so far have been with wire services and news agencies[19], which are already structured as feeds: short, fact-based bulletins that can be chopped up, recombined, and pushed into other products. It is much harder to imagine a rich, distinctive outlet turning itself into an API of atomised facts without losing its identity. Yet people may still find a generic AI overview 'good enough' for most purposes. This raises profound questions about what, if anything, distinguishes the value of a news organisation in such a world where AI agents will potentially deliver news

[18] See e.g. Stockwell, S. (2025, November 17). *From deepfake scams to poisoned chatbots: AI and election security in 2025*. Centre for Emerging Technology and Security (CETaS), The Alan Turing Institute. https://cetas.turing.ac.uk/publications/deepfake-scams-poisoned-chatbots
[19] https://petebrown.quarto.pub/pnp-ai-partnerships/

directly. One answer may lie in community. Participants highlighted that shared readership and recognisable editorial voices remain vital anchors for shared debate and dialogue. When you know that a particular outlet has published a specific investigation or made a particular mistake, that anchors a shared conversation. Community does not require comments sections; it can be as simple as thousands of people having read the same article and being able to talk about it.

This is also where healthy depersonalisation comes in. Speakers at the event argued that group-level 'bubbles' can still sustain common reference points in a way one-to-one personalisation cannot. One participant suggested that a single news site is in itself a kind of bubble, but it is a group-level bubble. Everyone who chooses to enter it sees roughly the same front page and the same editorial judgments about what matters. Personalisation can sensibly tweak around the edges, for example, not resurfacing stories you have already read or highlighting more local pieces if it knows the region where you live. But broader one-to-one personalisation, risks dissolving the shared space that democracy requires.

Newsrooms are having to adapt their internal practices as well with AI systems and AI agents increasingly common. Early on, some organisations tried to draw sharp lines: for example, any AI use for reader-facing content required prior approval from an editor, justified in terms of written principles. However, within a couple of years, this approach became untenable as AI has, in many cases, become too deeply embedded in everyday tools. During the event, newsroom leaders described a shift towards treating AI as a commonplace tool that nevertheless requires strong norms of accountability. The more realistic stance, one participant suggested, is to treat AI – and in future AI agents – the way Wikipedia was treated a generation ago: a tool that can be used, but one that in no way absolves journalists of the personal responsibility for everything they publish. Accountability and responsibility become a matter of agency; that is, journalists have the agency to use powerful tools, but they cannot outsource their judgment to them. As such, they cannot uncritically copy sentences or claims; they must check, contextualise, and rewrite.

This shift brings into relief how little training many journalists have had about what these systems are, how they work, and where they are reliable or dangerous. News organisations have often been slow to educate and create structured spaces for experimentation. Participants repeatedly noted a parallel between these professional gaps and the challenges faced by ordinary citizens navigating AI-mediated information. Yet the same vulnerabilities that threaten professionals – over-reliance, misplaced trust, failure to check sources – also threaten ordinary citizens. If journalists struggle to use AI wisely, what about the average voter with no institutional support?

Despite the clear threats speakers also emphasised the importance of keeping more constructive uses in sight amid the wider anxieties. One such use is simply the ability to ask an AI chatbot or AI agents 'stupid' questions without any social cost. Many people,

including journalists, hesitate to admit gaps in their knowledge, and such systems can be remarkable tutors for basic background, for instance, explaining constitutional structures, unpacking jargon, and providing context about historical events. Another is the capacity to use these technologies as deliberative tools that can help people understand opposing perspectives.[20] There is promising work on systems that help surface areas of emerging consensus or identify statements that people across divides can agree on. In theory, an AI moderator could be both more scalable and more neutral than human facilitators, who inevitably bring their own biases to the table.

That said, during the event, participants raised difficult questions about who sets the goals for such ostensibly 'neutral' systems. Who designs the prompts? Who chooses the training data? Who tunes the objectives? A system designed to bridge divides could just as easily be repurposed to manipulate people. The same technologies that can guide people toward mutual understanding could also nudge them toward particular outcomes. Ultimately, a machine can be many things, and we need to think carefully about what we want our machines to be.

Governance Challenges All the Way

All of these threads – personalisation, provenance, poisoning, persuasion, professional practice – converge. As participants summed up in the closing discussion, many of the most promising interventions are essentially design and governance choices.

First, participants argued that we need to build systems that make sources and time visible by default, not hidden behind icons or secondary clicks. Models should habitually say 'according to X, as of Y' when stating contested or time-sensitive facts. Second, they argued that we should treat human-authored, authoritative content (as well as the institutions producing it) as a kind of critical infrastructure, not just another data source that agents can draw on. That may mean technical standards for signing and labelling, legal protections for news organisations and other information provider that opt out of unrestricted scraping, and funding and licensing models that reward depth and verification rather than sheer volume and provide funding to organisations whose data gets used for the purposed of AI training and at inference time. Third, we should rethink personalisation itself, so the view in the room. It is not inherently bad that different citizens care about different issues, or that someone who already understands a country's political system doesn't need a 'Civics 101' course every time. The problem is when engagement-optimised personalisation completely swamps other values. Speakers argued that more 'civic-aware' personalisation regimes in AI agents and citizen-facing AI systems more generally are both possible and urgently needed. Better personalisation would explicitly incorporate goals like diversity of exposure, balance of viewpoints, and a minimum common core of 'things everyone sees.' It would also give users real awareness and control over how their feeds are being shaped, rather than hiding those choices in obscure settings that providers have no incentive to foreground.

[20] Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., ... & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science, 386*(6719). DOI: 10.1126/science.adq2852

Finally, as one participant put it, news organisations in particular might want to consider to stop doing what they do not, in their bones, care about. In a world where platforms and agents can churn out infinite generic updates, the comparative advantage of journalism might lie precisely in the opposite direction: in the hard choices about what matters enough to investigate, explain, stand over, and, if necessary, correct in public. That is not a technological function. It is an editorial one – and if democracy is to survive the age of agents, we will need more of it, not less. The final panel of the event concluded on this note: that strengthening editorial judgment and public-interest journalism is inseparable from managing the rise of AI agents as gateways to information.

CONCLUSION

As agentic AI continues makes its way into critical sectors and public life, its risks, benefits, and societal implications will only become more tangible. Managing the tension between usefulness and risk will require not only technical guardrails but also robust oversight, sector-specific safeguards, and mechanisms to maintain transparency and accountability. The symposium's aim was not to try resolving any of these debates but to provide clarity about the current state of affairs and surface the uncertainties and questions most in need of further research, experimentation, and scrutiny.

In closing this report, we would like to draw attention to some of the questions in need of further discussion, scrutiny, and research that emerged during and after the conversations at the symposium.

Defining and Understanding Agentic AI
What descriptive or evaluative approaches can help us map the spectrum of agentic AI autonomy in ways that inform governance? How can we assess and manage the tension between usefulness and risk in agentic AI? How can we productively conceptualise and apply the notion of 'human-in-the-loop' when it comes to agentic AI? What other principles might help inform responsible uses of agentic AI and how do these vary across domains? How can we better understand the limits of 'chain of thought' reasoning in agentic AI, and what complementary techniques might expand its safe capabilities?

Governance and Transparency

- In what domains is agentic AI already operating largely outside public visibility, and where should we increase transparency around these uses?
- What balance between specificity and flexibility is needed when regulating agentic AI across different sectors?
- How do we systematically identify, quantify, and mitigate risks (new, systemic, and incremental) that agentic AI introduces?
- What tools can we develop and deploy to verify the provenance of AI-generated content (including text) at scale? What cues should be communicated to the public and how is this most effectively done?
- What lessons can be drawn from the regulation of previous disruptive technologies for agentic AI?

Sectoral and Workforce Impacts

- What impact will agentic AI have on workforces across sectors? Where are the biggest risks and opportunities?
- How should organisations and governments plan for reskilling, redeploying, and preserving human agency?
- What systemic risks arise from cross-sector

reliance on a small number of foundation models, and how can technical and governance strategies diversify the supply chain?
- How can we better understand and explain the problems arising in AI–human interaction? Are there domains where this collaboration improve outcomes, and why?

Information Ecosystems, Media, and Democracy

- What democratic values do we want our AI models to have or support? How do we operationalise these values when designing and deploying agentic AI?
- What concrete actions can help ameliorate the societal risks of public reliance on AI agents for information and news (e.g. polarisation, misinformation, manipulation, misperception of public opinion, etc.)? What role can we realistically envision for literacy initiatives, regulation, and beyond?
- How might the rollout of agentic AI affect information ecosystems differently across countries with distinct regulatory frameworks, political and media environments, and levels of openness to technology?
- What lessons from the social media era of information intermediation can be applied to agentic AI to safeguard information quality and integrity?
- What role can and should journalism play in an information ecosystem mediated by agentic AI?

Knowledge, Institutions, and Research

- What role should universities and academic research play in auditing agentic AI and educating the public?
- What kinds of collaboration might help improve the identification of risks and threats arising due to agentic AI?
- How can institutional forms of knowledge production avoid being left behind by industry and public uptake of agentic AI? How do we reconcile the much-needed debate and reflection with the rapid pace of innovation?
- What collaborative approaches can bridge industry, academia, and civil society to monitor AI developments and ensure diverse perspectives inform governance?

BIOGRAPHIES

Dr Amy Ross Arguedas *(Rapporteur)*

Amy is a Postdoctoral Research Fellow at the Reuters Institute for the Study of Journalism (RISJ), where she works on the Digital News Report. Her current research focuses on news audiences and the relationship between journalism and new technologies. She was formerly a member of the RISJ trust in news project, which extensively studies trust from an audience-centric perspective. She holds a PhD and an MA from the Media, Technology, and Society program at Northwestern University and a BSc in Communication Studies with a concentration in Journalism from the Universidad de Costa Rica. Before pursuing her doctorate, she worked as a journalist for five years at the Costa Rican newspaper, *La Nación*.

Dr Felix M. Simon *(Principal Investigator)*

Felix is the Research Fellow in AI, Information, and News at the Reuters Institute for the Study of Journalism (RISJ), a Research Associate at the Oxford Internet Institute (OII), and a Junior Research Fellow at Corpus Christi College, all at the University of Oxford. His research looks at the implications of a changing news and information environment for democratic discourse and the functioning of democracy. Since 2019, his work has focused on various aspects of AI in news and the public sphere, with a special emphasis on its use and reception, the shifting power dynamics between the news and the technology sector, and AI's role in misinformation and democracy. His research and commentary frequently appear in international media, among others, in *The New York Times*, *The Guardian*, *Politico*, and the *Financial Times*. He regularly writes and comments on technology, media, and politics for various international outlets, including as a monthly columnist for Nikkei and advises media organisations and companies on AI. Felix holds a DPhil and MSc from the University of Oxford's Internet Institute and a BA in Film and Media Studies from Goethe-University Frankfurt. He is a Faculty Associate at the University of Wisconsin-Madison's Public Tech Media Lab and affiliated with Columbia University's Tow Center for Digital Journalism.

Dr Linda Eggert *(Principal Investigator)*

Linda is an Assistant Professor in the Department of Philosophy at Stanford University. Before, she was an Early Career Fellow in Philosophy at Balliol College and the Institute for Ethics in AI at the University of Oxford. Linda works primarily in moral and political philosophy. Most of her work concerns issues in normative and practical ethics, and theories of justice. At the Institute for Ethics in AI, her work focused on the relationship between human rights, democracy, and AI. Currently, she is especially preoccupied with the ethics of delegating to AI, autonomous

weapon systems, and the right to a human decision. Linda held fellowships at the McCoy Center for Ethics in Society at Stanford University, the Edmond and Lily Safra Center for Ethics at Harvard University, and the Carr Center for Human Rights Policy at Harvard's Kennedy School of Government, where she was among the inaugural cohort of Technology & Human Rights fellows. Linda holds a DPhil and MPhil in Political Theory from Oxford and a BA in Humanities, the Arts, and Social Thought from Bard College Berlin.

LIST OF ATTENDEES

The main panellists included:

- **Professor Rahul Santhanam**, Professor of Computer Science at Oxford University and a Tutorial Fellow at Magdalen College
- **Jonathan Rystrøm**, DPhil student in Social Data Science at the Oxford Internet Institute (OII)
- **Gill Whitehead,** Executive Director at NatWest Group plc, Informa plc, the British Olympic Association, and Chair of the next Women's Rugby World Cup (England 2025). Visiting Policy Fellow, Oxford Internet Institute (OII). Former Group Director, Online Safety, at Ofcom
- **Dr Adam Mahdi,** Group Lead, Reasoning with Machines AI Lab (RML) and Departmental Lecturer, Oxford Internet Institute (OII). Fellow of Wolfson College, University of Oxford
- **Professor Scott Hale,** Associate Professor and Senior Research Fellow at the Oxford Internet Institute (OII) and a Fellow of the Alan Turing Institute
- **Chris Moran,** Head of Editorial Innovation and the Editorial Lead on Generative AI at The Guardian
- **Professor Zeynep Pamuk,** Associate Professor of Contemporary Political Theory and Professorial Fellow at Nuffield College, University of Oxford

The attendees were from the following organisations: The Reuters Institute for the Study of Journalism, the Oxford Internet Institute, the Financial Times, the Faculty of Law at Oxford University, Said Business School, Stanford University, BOOM, Scroll India, the Associated Press, the BBC, City University of London, Big Brother Watch, the Financial Conduct Authority, Harvard Kennedy School of Government, El Diario, the Institute for Ethics in AI, TechUK, PoolDeck and a number of people in independent capacity.

ACKNOWLEDGMENTS

We would like to express our gratitude to the team at the Balliol Interdisciplinary Institute (BII), especially Elinor Richardson, and Balliol College for their generous funding, which enabled us to carry out this project. We are also grateful to the events team at Balliol College, thanks to whom everything went smoothly on the day of the event. We are indebted to Dr Amy Ross Arguedas, both for her extensive and exceptional notetaking during the symposium, and for her help in putting this detailed report together. We are also grateful to Daniel Patiño for the terrific graphic design and the layout. Finally, we would like to express our sincere thanks to all the panellists and participants, whose valuable perspectives and insights shaped the symposium and, subsequently, this summary report. Their input and participation were instrumental in informing the findings and recommendations presented herein. The usual disclaimers apply.

Oxford, December 2025