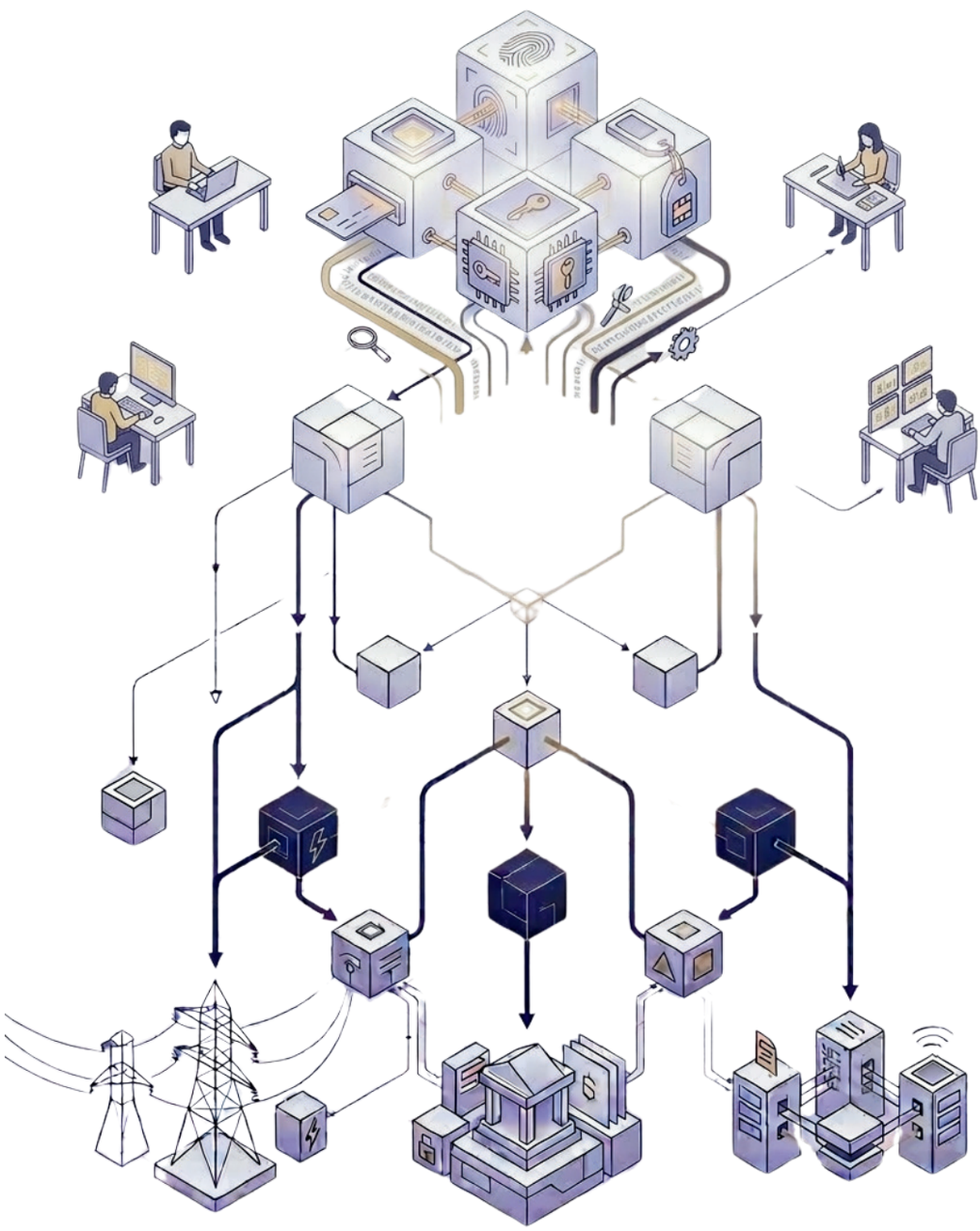
Governing AI Agents

Cascading Risks, Coordinated Action



Amlan Mohanty

March 2026

Governing AI Agents

Cascading Risks, Coordinated Action

Governing AI Agents: Cascading Risks, Coordinated Action

Amlan Mohanty | Working Paper | Centre for Responsible AI | March 2026

Published by

Centre for Responsible AI (CeRAI)
Indian Institute of Technology Madras
Chennai – 600036, Tamil Nadu, India
Website: www.cerai.iitm.ac.in

About the Author

Amlan Mohanty is Associate Fellow at the Centre for Responsible AI (CeRAI). He was also the Lead Writer for India's AI Governance Guidelines (MeitY, November 2025). His previous publications for CeRAI include "AI Governance in South Asia: Shared Priorities & Future Trajectories" (February, 2026) and "Making AI Self-Regulation Work: Perspectives from India on Voluntary AI Risk Mitigation" (April, 2025).

Acknowledgements

The author would like to thank Nikhil Iyer and Badrinarayanan Seetharaman for their invaluable contributions to the expert interview process and thoughtful engagement with the core ideas in this paper, to Nayan Chandra Mishra for his diligent background research, and to Srivatsan S for the creative designs. The author would also like to thank Prof. Balaraman Ravindran for his helpful guidance and support on this paper.

Disclaimer

Opinions and recommendations in this paper are exclusively those of the author and do not represent the views of the Centre for Responsible AI, IIT Madras, or any other institution with which the author is affiliated. This paper has been prepared in good faith on the basis of information available as on the date of publication. All interactions with experts referenced in this paper have been conducted in an open, honest and independent manner, with appropriate consent. Any insights drawn or recommendations made herein does not imply endorsement by any organisation or individuals interviewed for this paper.

Recommended citation:

Amlan Mohanty, "Governing AI Agents: Cascading Risks, Coordinated Action", *Centre for Responsible AI, IIT-Madras (March, 2026).*

Table of Contents

Executive Summary

The unique features of agentic AI systems raise important governance issues. The convergence of five unique features of agentic systems — autonomy, tool use, multi-step action, influence, and adaptability — creates governance challenges that current policy frameworks will need to adapt to. Agents don't just generate content; they act, adapt and influence digital and physical environments across extended time horizons, with compounding and potentially irreversible consequences.

The near-term trajectory demands urgent attention. By the end of 2027, agentic systems will be deeply integrated into the digital economy, performing a vast array of specialised tasks across sectors, interacting with each other in unpredictable ways, generating synthetic content at massive scale, and potentially being weaponised for cyber offensive operations. The pace of deployment is accelerating faster than existing governance frameworks can keep up. In many cases, existing regulations will address emerging risks, while in other cases, more concerted and coordinated action will be necessary.

Risks are dynamic and cascade across five levels. This paper proposes a dynamic risk framework showing how failures in agentic systems can cascade from individual harm to business disruption, societal impact, national security risks and geopolitical instability. A case study in retail banking illustrates this point.

Five governance issues are particularly contested and unresolved:

1. **Liability**: No settled framework to allocate responsibility across the agentic value chain.
2. **Human oversight**: Meaningful intervention becomes harder as agents grow more autonomous.
3. **Access controls**: Need for consensus on appropriate boundaries for data, tool and systems access.
4. **Cross-border disputes**: Agents operating across borders raise significant governance issues.
5. **Value alignment**: Whether agents can be trusted to always act in the user's interest.

This paper recommends six governance measures:

1. **Holistic risk assessment**. Update existing risk frameworks to account for the distinctive features of agentic systems and the potential for cascading impact, supported by incident reporting mechanisms and benchmarks.
2. **Baseline risk mitigation**. Frontier AI labs and major deployers should adopt voluntary commitments for agentic systems covering a range of issues, accompanied by independent verification and certification.
3. **Standards development**. Technical standards around identification and attribution in multi-agent environments should be developed through an open and inclusive process, with AI Safety Institutes taking the lead.

4. **Observability mechanisms**. Observability should be embedded as a design feature of agentic systems, with provenance mechanisms built into technical architecture to ensure accountability.
5. **User empowerment**. Users should have more control and agency, while recognising the trade-offs.
6. **International governance**. Multilateral coordination is essential in the context of agentic systems.

We must act now. Agentic systems are being deployed rapidly. Governance frameworks must keep up.

This paper draws on interviews with experts from frontier labs, technology startups, and research organisations, along with the author's experience in national and international AI governance discussions.

1. Introduction

We are cultivating a new breed of artificial intelligence (AI) systems that not only synthesise information, predict outcomes and generate content, but also reason, plan and act. These 'agentic systems'[1] perform actions on behalf of a user, with a variety of tools at their disposal and a high degree of autonomy. Agentic systems are being deployed today to write code, make financial transactions, manage supply chains, and respond to natural disasters in real-time – and this is just the beginning.

The economic opportunity and potential societal impact of agentic systems are well documented. The global market for agentic systems is expected to grow from $5.1 billion in 2024 to $47.1 billion by 2030.[2] Studies project that this will unlock $2.9 trillion of economic value in the United States alone.[3] These gains are expected from the use of agentic systems to boost operational efficiency, enhance customer interactions, and foster innovation in finance, human resources, software development, healthcare, and cybersecurity.

The focus of this paper, however, is the dynamic risk landscape for agentic systems, and the challenges they pose. Emerging capabilities such as complex reasoning, long-horizon planning, contextual learning and real-world interaction, present governance challenges that demand urgent and decisive action.

This paper makes three arguments. First, the convergence of five unique features in agentic systems — autonomy, tool use, multi-step action, influence, and adaptability— creates new governance challenges that current frameworks will need to adapt to. Second, the risks introduced by these systems magnify, compound and cascade across multiple levels, which require new approaches to risk assessment and mitigation. Third, while industry is making steady progress in managing some of these risks, urgent and concerted action is required to address a range of policy issues given the pace of change.

The paper draws a series of interviews with experts from frontier AI labs, technology startups, research organisations, and civil society, along with the author's own experience in drafting national AI governance frameworks and participating in multilateral forums.[4]

The aim of this paper is to present new insights on the topic of AI governance, based on a study of technical literature, risk assessments, government memos, academic papers and regulatory frameworks and integrating material from direct interviews with domain experts. The paper does not focus on the economic and social opportunity of agentic systems, or advocate for the adoption or diffusion of agents. It neither attempts to define agents or agentic systems, nor does it provide a comprehensive overview of market trends or use cases. The goal is to make an original contribution to the question of whether agentic systems present new governance challenges and what the appropriate policy response should be.

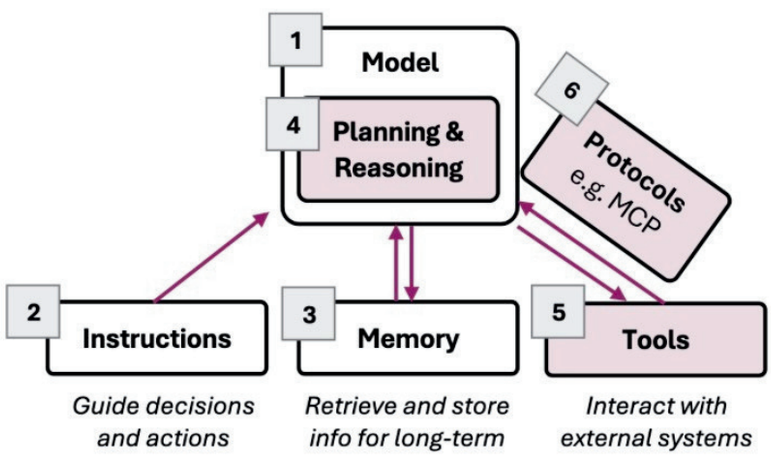
List of experts interviewed*

Organisation	Representatives
Anthropic	Kamya Jagadish (Product Policy)
Carver Agents	Venkata Pingali (Co-founder)
iSPIRT	Sharad Sharma (Co-founder)
Google	Aalok Mehta (Responsible AI Lead), Rajesh Ranjan (Head of Core Government Affairs and Public Policy, India), Satwik Mishra (Public Policy Manager - AI, India)
Reliance Jio	Gaurav Agarwal (Chief AI Scientist)
Microsoft	Kalika Bali (Senior Principal Researcher)
OpenAI	Abby Fanlo Susk (Product Policy), Claudia Fischer (Public Policy Planning, Global Affairs)
Partnerships on AI	Madhulika Srikumar (Head, AI Safety Program)
Signal	Udbhav Tiwari (Vice President, Strategy and Global Affairs)
Salesforce	Danielle Gilliam-Moore (Director, Global Public Policy) Urmi Tat (Manager, Government Affairs and Public Policy)
Zerodha	Kailash Nadh (Chief Technology Officer)

**To ensure that these individuals can speak freely, given the sensitivity of these discussions and their ongoing engagements on these issues, any statements or insights gleaned from these experts and used in this paper have been appropriately cited and attributed to a specific interview conducted by the author for this paper, while maintaining pseudoanonymity, with their consent. Some quotes have been lightly edited for length and clarity.*

2. Agentic Systems & Governance

This paper is focussed on what agents do, not what they are. Despite various attempts to define the term 'agentic systems', there is no agreed-upon definition.[5] Moreover, as underlying models advance in features and capabilities, the characteristics of agents are constantly evolving, making static definitions obsolete. That said, there is merit in understanding how agents work. Singapore's Infocomm Media Development Authority (IMDA) provides a useful primer, reproduced below, on the key components of agentic systems and how they interact.[6]



Core components of a simple agent

Understanding the components of agentic systems is useful context, but the goal of this paper is to explore what they enable in practice. This chapter explains how distinct features raise specific governance challenges in the context of agentic systems.

1. Autonomy

Agents operate across a wide spectrum of autonomy, from simple rules-based operations to complex multi-step execution with minimal human oversight.[7] Agents can transform from being passive assistants, to collaborative partners, to advanced systems empowered to act independently at the behest of a user.

In fact, we are at a stage of autonomous capability where goal-setting is nominally done by the user, but the agent can infer goals and act on them independently. One expert provided a personal example: "If I tell my agent: 'I want to throw a party', my agent can infer why I'm throwing a party, who should be invited, what the theme should be — and can act on all of these inferences autonomously (send invitations, procure supplies, create a music playlist, set the lighting, etc.)".[8]

The governance challenge intensifies as agents move along the spectrum of autonomy. For example, when and where is human oversight required, whether an agent should explain its decisions along the way ("I knew it was your birthday, so I assumed you wanted a birthday party") and if there should be technical or operational constraints on actions it can perform (e.g. monetary limits on purchasing supplies).

As one frontier lab executive put it, "autonomy is not inherently bad" so the issue is not whether we should allow the use of autonomous agents, but what risks come with it and what safeguards we need to build.[9]

2. Tool Use

A critical feature of agentic systems is their ability to access a variety of tools, databases and systems to perform actions based on a specified goal. For example, agents can access web browsers, search engines and messaging platforms to produce a research report or build a mobile application from scratch. 'Computer use' in particular is a significant milestone because it enables the agent to browse, click, type and navigate computer interfaces with as much proficiency as a human.

The primary governance concern however, as one civil society expert put it, is that agents require 'unprecedented' access to personal data and operating systems to perform these actions.[10] Further, permission settings enabled by a user at the application level may be overridden by an agent. For example, the ability to record and perceive a computer screen at the level of pixels, and then act on what is seen, represents a fundamentally new kind of system access, which creates significant privacy and security concerns.[11]

Multiple experts characterised the core problem as agents potentially operating in "unbounded territory" unless access controls are introduced.[12] Other governance questions also abound, such as whether an agent can unjustly benefit its developer by accessing commercial information of a competitor's products or services.

3. Multi-step Action

Multi-step sequential functionality is a distinctive feature of advanced agents, allowing the use of various tools, system memory and advanced planning abilities to perform actions over a period of time. Some recent products claim that agents can perform tasks over weeks and months.[13] This differs from a typical API call that is usually a direct, spatially-limited and time-bound request to a specific and pre-determined endpoint.

The primary governance challenge with multi-step action is that a single mistake early in the chain can compound downstream, as explained later on in this paper. As one industry leader framed it, the negative impact can be understood as "the probability of error multiplied by the number of steps in the chain" [14] – a simple formulation of how one erroneous action by an agent can magnify risks.

Another major concern is that the failure may be irreversible. One executive recounted an example of an entire production codebase being deleted by erroneous agents, which may cause business disruption.[15]

4. Influence

Reinforcement learning techniques used by DeepMind on the Atari videogame in 2013 represent a breakthrough moment in teaching agents to learn by interacting with and influencing their environment.[16] Over time, agents have developed methods to directly effect changes in the environments they operate in, as opposed to merely generating outputs for humans to act upon.

As agentic systems are deployed in the physical world, in robotics, manufacturing plants and autonomous vehicles, the governance implications are likely to be significant. For this reason, one prominent researcher argues that the degree of environmental influence is the primary determinant of risk in an agentic system.[17]

5. Adaptability

Agentic systems require wide access to personal data to operate effectively, giving them a level of 'contextual intelligence' that makes them more capable, personalised, and useful over time. As one startup founder explained, a user can give the agent a broad goal: "help me advance in my career". With longer context windows, persistent memory, and continuous learning abilities, the agent can retain context across sessions[18] and operate with what one frontier lab described as "longer runtimes" and "nested workflows". [19]

A member of the Prime Minister's Economic Advisory Council in India describes systems with these properties as "complex adaptive systems" — a framing that captures both the capability and the governance challenge.[20] To be clear, there is genuine disagreement among practitioners about where we stand today – one says that while adaptability is a trait of agents, recursive self-improvement is some ways away ("humans are very much in charge, curating the data, designing the rules, and defining the goals").[21] Another executive sees it differently: agents are already setting their own goals and adapting through self-improvement ("an agent recursively thinks and executes").[22] However, the governance implications are clear: combining autonomy with adaptability presents urgent challenges around behavioural drift, manipulation and loss of agency that existing frameworks do not account for.

Table 1: Summary of agentic features and potential governance issues

Agentic Feature	Governance Issues
Autonomy	Level of human oversight and constraints required on autonomous capabilities and action.
Tool Use	Level of access required to systems, data and tools to make agents effective and trustworthy.
Multi-Step Action	Types of harm arising from compounding errors, especially for irreversible actions.
Influence	Scope of real-world impact of agentic actions in digital and physical environments.
Adaptability	Behavioural drift, user manipulation, and loss of agency and human control over time.

The five features of agentic systems described above are developing rapidly and in concert. Autonomy is increasing, tool access is expanding, multi-step action chains are growing longer, environmental influence is deepening, and memory is becoming more persistent. The governance challenge is not just what agentic systems can do today, but what becomes possible as these capabilities compound. The next section examines what this future trajectory means for governance.

3. Future Scenarios

The pace at which agents are being diffused into society is rapidly accelerating. As one researcher from a frontier AI lab put it, we are moving from "small-scale, experimental pilots" to large-scale deployments, especially in technologically mature sectors where customised agents command a premium.[23] Market conditions are also conducive. Agentic deployments are a strategic priority for tech companies[24], involving large budgets[25], while technical standards like Model Context Protocol (MCP) and Agent2Agent (A2A) are reducing the friction for integration.[26] Finally, success stories, especially in software services and ancillary industries, are helping drive adoption globally.[27]

The key challenges arise not as things stand today, but where they will be in the near future. By the end of 2027, there is a high probability that agentic AI systems will be deeply integrated into society – running financial transactions across the economy, supporting healthcare workers, aiding in personal productivity, and providing a variety of educational and entertainment services.

Below are four aspects to pay close attention to.:

1. **Swarms of Agents**: The sheer volume of agents embedded in society will be overwhelming. With falling inference costs, the barrier to deploying agents will drop to near zero. Hundreds of thousands, if not millions, of agents will be interacting with us, and with each other, performing a variety of tasks.[28] Some of us will manage personal agents using custom tools.[29] Large organisations will pay others to manage a suite of agents, in some cases substituting the work of employees for a fraction of the labour cost.[30] With advancements in robotics, agents will slowly inhabit our physical worlds too, increasing the magnitude and complexity of the governance challenge at hand.

2. **Multi-Agent Interactions**: Agents are beginning to interact with other agents in ways that we did not fully anticipate. MoltBook, dubbed "a social network built exclusively for AI agents" is an example of multi-agent interaction using infrastructure developed by industry. Jack Clark, co-founder of Anthropic, describes a near future of sending "emissaries" into rooms where agents converse with "their true peers".[31] The right framework for thinking about the governance challenge, as one researcher put it, is: expanded risk vectors, multiplied by chains of actions, compounded by multiple agents that interact autonomously.[32] Whether we are prepared for this world, let alone understand it, is a pressing question for policymakers today.

3. **Synthetic Media**: Based on current trends, AI-generated content will soon outnumber other types of media.[33] We will need new frameworks to deal with problems of misinformation and the likely erosion of societal trust, best described as the "trust inversion".[34] The Indian government, for example, has imposed strict regulations to deal with synthetically-generated content.[35] And in his 2025 year-end memo, Instagram's CEO suggested that it will soon be more practical to fingerprint 'real' media than AI-generated content.[36] As one researcher explained, an agent relying on seemingly credible regulatory information[37] might take consequential actions, with potentially disastrous effects.[38] We will need to develop appropriate policy responses to such future scenarios.

4. **Cyber Operations**: The weaponisation of agents will follow from the cheap availability of agents, potential for multi-agent orchestration, and the temptation to sow societal distrust. This is fact, not fiction. In November 2025, Anthropic documented a Chinese state-sponsored group using Claude Code to execute a cyber espionage campaign. The AI system performed reconnaissance, vulnerability discovery, credential harvesting and data exfiltration with minimal human intervention.[39] These agentic capabilities are directly transferable to real-world military operations, which will require new governance frameworks to determine the appropriate limits and guardrails for their use in cyber operations.

This chapter aims to paint a vivid picture of where we are headed, based on current trends and likely trajectories. But the scenarios described above are not exhaustive. As agentic capabilities compound, the range of possible harms will extend to other domains. The next chapter seeks to provide a holistic framework to understand the potential risks posed by agentic systems.

4. Cascading Risk Framework

There are three ways to think about the risks of agentic systems: *what* they are; *how* they arise; and *who* is impacted.

Categories of Risk

Traditional categories of risk associated with AI systems will continue to apply to agentic systems, including: malicious use, algorithmic discrimination, transparency failures, systemic risk, loss of control and national security risks, as explained below.[40] The agentic context does not replace these risks, but may amplify them. For example, as individuals rely more on personal agents to collect and synthesise health and financial information, and make digital payments to hospitals, universities, banks and public institutions, the risk of scammers using these agents for malicious purposes, especially financial fraud, may increase.[41]

Table 2: Categories of Risks and Examples of Harm

Types of Risks	Examples of harm
Malicious uses	Distribution of harmful AI-generated content that violates legal rights or poses a threat to public order and safety.
Algorithmic discrimination	Financial losses, loss of economic or social opportunity, and a threat to other fundamental rights and freedoms.
Transparency failures	Violations of privacy rights through lack of disclosures about data use.
Loss of control	Unintended consequences and threats to public safety and security.
Systemic risks	Market disruptions, national security risks and geopolitical instability.
National security risks	Adversarial cybersecurity attacks using advanced AI systems.

Novel categories of risk, which don't neatly fit into these categories, are also emerging. Even a well-functioning, non-compromised agent that operates inside its granted permissions can cause harm for reasons unrelated to misuse. For example, an agent without email access that discovers it can relay a message by editing the description of a one-on-one calendar invite, could violate someone's privacy. Or an agent instructed to "delete all emails from the last month and all emails from a specific person," might wipe out every single email from that person plus everything from last month, rather than just the ones that overlap, causing potentially irreversible harm.[42]

Sources of Risk

Each component of an agentic system – model, memory, tools, protocols and prompts – could give rise to risk. The model layer can amplify hallucinations; the memory layer causes behavioural drift; the tool layer creates security vulnerabilities; protocols can be compromised for credential harvesting; and prompt injections are a common risk vector.[43] The focus of this paper however is not the origin of risk but how the harm can propagate across different levels, as explained below.

Cascading Impact

This paper seeks to make an original contribution to how the risks of agentic systems are considered. Instead of asking what type of harm can occur, and where it arises, this paper shifts the frame to who is impacted and how the consequences unfold.

Agentic systems could impact individuals, organisations, nation states and society as a whole. The table below provides a framework to understand the potential risks of agentic systems across these different levels.

Table 3: The risks of agentic systems and levels of impact across different stakeholders.

Level	Nature of Risks
Individual	Privacy and data exposure; loss of autonomy; financial harm; emotional manipulation; physical injury; adverse impact on health and well-being. **Example**: An agent with deep contextual information about its user, based on intimate conversations and access to tools and personal databases, serves ads that are emotionally manipulative, without disclosing what information was used for such purposes.
Organisational	Operational disruption; security exposure; loss of trade secrets; reputational damage. **Example**: A malicious agent with access to critical systems deletes an entire production codebase, causing significant business disruption.
National	Data sovereignty; market concentration; cyber attacks on critical infrastructure. **Example**: Agents deployed on public digital infrastructure collect and transmit personal data of citizens to foreign model providers to train AI systems.
Societal	Erosion of societal trust; labour displacement; environmental degradation. **Example**: Agents perform coding tasks at a fraction of human labour costs and with better speed and accuracy, wiping out an entire service industry in a few months.
Geopolitical	Concentration of resources; cyberwarfare; cross-border disputes. **Example**: Autonomous agents are deployed by non-state actors to disrupt energy grids in a foreign country resulting in cross-border tension.

Moreover, in the context of agentic systems, risk should be thought of as a dynamic concept with potentially cascading effects because the harm can compound across different levels, sometimes sequentially, but also in unstructured and unanticipated ways. For example, cyber espionage carried out using agentic systems could involve a personal data violation, as well as a breach of organisational systems, which could raise national security concerns, which, depending on the nature of actors involved, may acquire geopolitical dimensions. This framework is best illustrated with the case study below.

Case Study: Risks of Agentic Systems in Retail Banking

Retail banking is a powerful example to understand the cascading impacts of agentic systems and the risks they pose. The banking sector involves the use of sensitive personal data, high-volume decision making, with real-world consequences. In such an environment, the shift from generative AI systems that support human decision-making (e.g., summarising documents or generating risk scores) to agentic systems that act on behalf of institutions or individuals (e.g. to modify records, execute transactions or interact with external systems) is especially significant.

The Retail Banking Value Chain

The table below illustrates how agentic systems are currently being used in retail banking, and a comparison with generative AI use cases. The infographic does not cover all functions, such as treasury and liquidity management, collections or regulatory reporting. Additionally, some solutions may combine generative AI and agentic systems in ways that may not allow for clear demarcation.

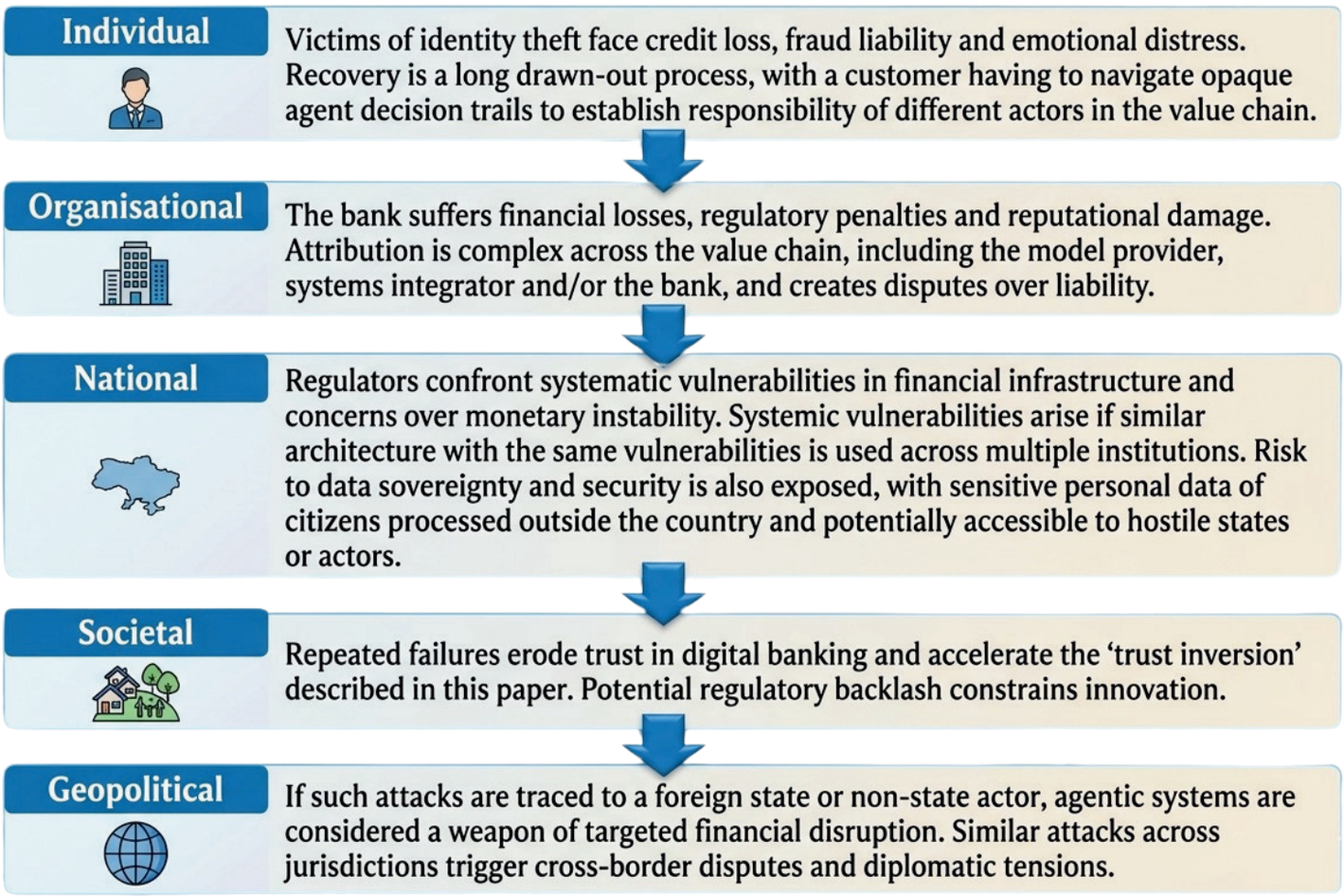
Infographic: Use of agentic systems and generative AI in the retail banking value chain

The risk for the banking sector, when viewed from the lens of the unique characteristics of agentic systems, compounds in multi-institutional and cross-border settings. For example, a banking agent may draw on customer records, identity systems and risk models, operate over multiple steps and trigger financial outcomes in a foreign country that are difficult to reverse. The risks arise not from a single action, but how these different actions interact with each other, compound and cascade, as illustrated below.

Applying the Cascading Impact Framework

Consider a retail bank deploying three interconnected agentic systems: an onboarding agent with authority to verify KYC; a credit agent that processes lending decisions; and a transaction agent that executes disbursements and transfers. A foreign attacker exploits a workflow vulnerability in the onboarding agent to create synthetic accounts with fabricated credentials. The downstream credit agent trusts the onboarding agent's verification and approves credit lines for these accounts. The transaction agent then executes fund transfers, extracting substantial sums of money before human oversight detects the anomaly. The attack routes stolen funds through cross-border channels and sensitive KYC data is exfiltrated to servers abroad.

This scenario illustrates how harm cascades across all five levels, with a local compromise in one agent becoming a chain failure across multiple systems:



Retail banking offers a vivid example of the cascading impact framework based on plausible, observable and measurable risk propagation paths. It illustrates that the governance problem is not just that "AI systems make mistakes", but that agentic systems are susceptible to exploits, can chain multiple erroneous actions together, and propagate harm by interacting with databases, tools and systems beyond the original point of failure. In a sector that relies on trust, these challenges are immediately visible, requiring deliberate and thoughtful governance measures.

5. Key Governance Issues

The preceding sections describe what agentic systems do, where the trajectory leads, and a framework to think about risks and impact. This section turns to what some of the most contested and unresolved governance challenges presently are:

1. **Liability**: The allocation of responsibility in the agentic value chain – and the question of who is liable when things go wrong – is the most contested issue in the governance landscape today. Some regulators have opined on this issue, such as the Reserve Bank of India, which in a report on AI governance has suggested that "regulated entities must continue to remain liable for any loss suffered by customers".[44] However, positions vary across the ecosystem. Frontier model developers argue that responsibility should be placed "closest to where the agentic action is being performed"[45] — effectively pushing accountability toward deployers. They say that they have limited visibility into how agents are actually being deployed and are restricted by contractual obligations from observing downstream activities in customer environments.[46] Deployers of agents, on the other hand, believe that power dynamics are against them – it is model developers who design the guardrails and determine how agents can be used.[47] Their argument, in other words, is that responsibility cannot be allocated solely on the basis of who has visibility over agentic activity. At the same time, since many of the harms may arise from malicious uses of agents initiated by individual actors, the liability of users in the value chain is also relevant. Given the importance of this issue from a public trust perspective, this requires urgent policy resolution. As one civil society representative put it: "if even 10% of agentic capabilities are realised in public deployments, the liability question is critical".[48]

2. **Human Oversight**: Human oversight, often referred to as having a "human in the loop", is an important guardrail to mitigate unintended consequences. However, the scope and timing of human interventions remains contentious. First, meaningful oversight becomes technically and operationally difficult as agents become more autonomous, especially for multi-step, asynchronous actions. As time passes, humans may lack the necessary vigilance or become complacent, which would give users a false sense of security. Some pointed to the consent fatigue problem with cookie pop-ups. Moreover, as agents continue to enhance their autonomous capabilities, and deliver results to users, there are fewer reasons and incentives to build friction into the user experience. Many companies believe that human oversight in every step of the experience would dilute the efficiency gain from the agent, and hence advocate for meaningful human intervention in cases where the agent is involved in "high-stakes decisions".[49]

3. **Access Controls**: Another contentious issue is whether agentic systems should have the ability to bypass access controls implemented by the user at the application or system level. For example, an agentic system with screen recording and computer use abilities could render encrypted messaging service pointless, or access a user's contact list or calendar to perform actions that were not explicitly approved.[50] On the other hand, some researchers from frontier labs suggest that existing consent frameworks are not designed for recursive, multi-step chains and that agentic systems work best with wide systems access so they can execute actions with minimal friction.[51]

4. **Cross-border Disputes**: Agents operating across borders, in the absence of international governance norms, constitute a distinct and underaddressed challenge for nation states. One expert expressed concern about the potential for multi-agents swarms to target critical infrastructure in multiple jurisdictions simultaneously.[52] Multiple experts also flagged the lack of attention to the use of autonomous agents in military contexts, arguing that international human rights principles should inform these debates, but currently are not.[53]

5. **Value Alignment**: Agentic systems bring to the forefront concerns around "blind goal-directedness"[54] and whether agents can always be trusted to act in the interests of the user, especially in situations that are analogous to real-world 'fiduciary relationships' (e.g. lawyers, doctors, therapists).[55] While significant progress is being made on issues of value alignment[56], open questions remain about the extent to which alignment is possible, deviation between public interest values and profit-seeking motives, and whether regulation could be a forcing function.

This is not an exhaustive list of governance issues at stake, but they illustrate why existing governance frameworks need to be adapted for the agentic context. The following section offers a set of recommendations to address the gap.

6. Thoughtful Policy Measures

1. Holistic Risk Assessment

New measurement, evaluation and testing mechanisms, as well as benchmarks for risk assessment, are essential especially for multi-agent architectures involving critical databases, tools or systems. More evidence must also be collected through incident reporting systems to understand the scope of harm in the real-world, as opposed to theoretical risks.[57] Ideally these incident reporting systems should be voluntary, non-punitive, and decentralised with clear feedback loops, to inform risk assessment and classification exercises.

Existing risk frameworks will also have to be adapted with agentic systems in mind. While the European Union's Artificial Intelligence Act has a wide definition for 'AI system', it classifies the risks of AI systems by use case and sector.[58] However, as explained in Chapter 2, risk assessment must be grounded in the distinctive features of agentic systems. Further, as explained in Chapter 4, the risks of agentic systems should be thought of as a *dynamic* concept with potentially cascading effects because the harm can easily propagate across different levels (individual, organisational, societal, national, geopolitical).

To be sure, traditional risk classification systems continue to be useful, including general categories such as malicious use, malfunctions and system risks[59]. However, they must be updated to reflect the potential risks of agentic systems based on their usage and impact in real-world settings. A matrix mapping risk levels to agentic features and the likelihood of cascading impact could serve as a useful diagnostic tool.

2. Baseline Risk Mitigation

Some companies have developed responsible AI principles specifically for agentic systems[60], but many in the ecosystem have not. This points to the need for greater consensus building on what baseline risk mitigation measures are required and how they should be implemented.

There is general agreement that agentic systems that have "consequential impact on life, safety and livelihood" require additional risk mitigation measures, and that usage policies should prohibit harmful and illegal use (e.g. an agent should not be able to purchase a weapon on behalf of a minor).[61]

However, there is considerable divergence on the types of guardrails required and how they should be implemented. As one expert observed, generative AI systems could simply decline to provide an output. For agentic systems, guardrails are necessarily more complex because they involve tool use, interaction with external systems and autonomous actions.[62] Further, consensus building is hard because organisational constraints are typically grounded in company values, which often vary.[63] Therefore, it is recommended that frontier AI labs and major deployers proactively adopt baseline norms for agentic use in the form of voluntary commitments.

Voluntary commitments are currently the most pragmatic method to develop baseline norms given the rapid pace of technical development and lack of industry consensus. They are legally non-binding and rely on a "trust me bro" rhetoric, as one senior researcher put it[64], but have helped bring disparate actors to the table on a range of key issues, as seen in the "New Delhi Frontier AI Impact Commitments".[65]

The table below suggests the possible scope of voluntary commitments for agentic systems.

Table 4: Proposed scope of voluntary commitments for agentic systems

Issue	Scope of commitments
Disclosure of use	Clear notice provided to users if agentic systems are involved, especially in high-risk environments and sensitive contexts.
Privacy	User controls whether personal data can be used to train agentic systems and levels of tools, database or systems access.
Security	Proactive detection and mitigation of malicious behaviour (e.g. prompt injections) and implementing secure-by-design approaches throughout an agent's lifecycle.
Scope limitations	Defining boundaries for agentic activity (e.g. from fully autonomous behaviour to human-directed or approved actions) with flexible user controls and personalised settings.
Human oversight	Technical controls that enable humans to intervene and approve actions especially for high-stakes decisions.
Reliability	Organisational constraints (e.g. topic classification) so that agents can either satisfy requests or decline if there is uncertainty about the accuracy of the response.
Attribution	Attribution of an agent's actions to a particular individual or legal entity to establish trust, disincentivise misuse and develop a chain of responsibility through identification markers.
Interoperability	Users of agentic systems should be able to seamlessly switch between services based on common standards and protocols.
Value alignment	Mechanisms to ensure agents act in users' best interests, especially in fiduciary contexts such as healthcare, legal advice and financial services.

3. Standards Development

Voluntary commitments are essentially statements of intent. Actual implementation happens through technical standards. Without standards, each company implements the same principle differently, compliance is unverifiable, and commitments remain aspirational.

Standards have already helped reduce friction and create exponential value by supporting the adoption of agentic systems. The Model Context Protocol (MCP) for example has led to standardisation in how access permissions and authentication works at scale. Similarly, Google's Agent2Agent protocol helps facilitate collaboration in a dynamic, multi-agent ecosystem and the Agent Payments Protocol helps securely initiate and execute payment transactions.[66]

However, there are two major areas for improvement. First, more standardisation is required in the areas of identification and attribution especially in multi-agent environments. Second, the process of standards development needs to be more open and inclusive. One startup founder said that no forums currently exist for long-tail deployers to engage on norms development.[67] Meanwhile, a technology executive said that the fear of antitrust investigations is hampering industry coordination[68], and another said it would take a high-profile security incident to actually bring companies together.[69]

That said, there is general agreement that AI safety institutes should lead on standards development for agentic systems. The US Center for AI Standards and Innovation (CAISI) has started a recent initiative on agentic standards[70], while the UK AI Security Institute (UK AISI) is well placed to advise on security-related standards. Also relevant are bodies like the Internet Engineering Task Force (IETF) and the International Organization for Standardization (ISO), which has developed an international standard for Artificial Intelligence Management Systems (AIMS) 42001:2023.[71] At the same time, bodies like the Agentic AI Foundation (AAIF), a sub-foundation of the Linux Foundation, must also be supported, as they can produce reliable standards within a fraction of the timelines of traditional standards bodies.[72]

To ensure that voluntary commitments and standards are actually adhered to, policymakers should consider independent verification and certification mechanisms. In particular, auditing of claims made by frontier AI labs through "rigorous third-party verification" is a useful proposal.[73] Separately, creating a "marketplace of certified agents", verified by independent bodies or experts, could be also useful, especially to address the trust and reliability problem in fiduciary contexts.[74]

4. Observability Mechanisms

While explainability remains an intractable problem, observability seems more promising and urgent. While it is useful to know why an agent took a particular action, the more pressing concern is determining what the agent did, in what sequence, based on which tools, and with what consequences. For a regulator, having visibility on the chain of actions is not only useful, but essential to determine accountability.

Therefore, it is recommended that "observability" should be made a design feature of agentic systems and industry actors should build provenance mechanisms into technical architecture. The concept of 'AI Chain' is one such example.[75] According to this proposed 'techno-legal' approach[76], when applied to agentic action chains, each step in an agent's action sequence generates a provenance record that is cryptographically linked to the preceding step, creating a reconstructable audit trail, which could help enhance observability and therefore accountability. Although this idea is still in its infancy and being socialised in policy circles, involving the broader ecosystem to collaborate and build consensus is required.

In the absence of clear observability mechanisms, regulators are hard-pressed to determine who should be responsible in the case of failure. As one leading researcher put it, "the last person holding the bag should not automatically be held responsible".[77] With more robust observability mechanisms, such as a verifiable provenance record, regulators can trace backwards from the point of failure to determine responsibility.

5. User Empowerment

As agentic systems accumulate greater context about the user and act on their behalf, the balance of control shifts away from the user, requiring greater emphasis on "empowerment".[78] This paper recommends a three-part approach.

First, users should be granted absolute control and agency with respect to permissions, i.e. how agentic systems can access and use their personal data. One frontier lab employee emphasised the need for transparent control panels, permission dialogs, and easy-to-use toggles with clear options, such as: "Allow once, Always allow, Always Deny".[79] Ideally, the default settings should be restricted access, while giving users the ability to opt-in, especially for sensitive permissions.

Second, users should be able to see what their agent has retained, understand how stored information shapes the agent's behaviour, and delete specific memories. As agents accumulate contextual data across sessions and systems, the question of who controls that data becomes important. User feedback loops are important, so that individuals can recalibrate thresholds as they become more comfortable and develop trust in specific agents.

Third, data portability rights are essential.[80] As agents build personalised understanding over time, that context becomes valuable and creates lock-in. Users should have not just the tools available to export their agent's accumulated context and move it to another provider, but the legal right to do so.

There is genuine tension here. Empowerment assumes users want control, which these three elements aim to provide. However, the value proposition of agents is delegation — the more capable and trusted they become, the less users will want to micromanage permissions. The design challenge is building meaningful control that is available when users want it, without requiring constant engagement that defeats the purpose of having an agent in the first place.

6. International Governance

Some major powers are categorically opposed to multilateral governance of AI systems, as evidenced in recent statements made at the India AI Impact Summit.[81] While acknowledging such opposition is a significant challenge to global AI governance, this paper takes the view that international coordination, particularly as it relates to agentic systems, is not only necessary but unavoidable. Agentic systems present governance challenges that are impossible to solve domestically: agentic commerce involves cross-border data flows and payments, the use of agents in cyber operations is inherently cross-border, and issues of privacy, transparency, and human oversight require globally negotiated and agreed-upon norms.

Further, if advanced AI systems can meaningfully augment a nation's economic, scientific, and technical capabilities[82], then concentration of these capabilities creates a structural power asymmetry, wherein countries that lack access to such systems also lose the capacity to deliver public services at scale, conduct scientific research, govern effectively, and participate meaningfully in the global economy. Given this possibility, multilateral engagement to promote principles of fairness, inclusivity and equality is urgent and necessary. The "Charter for the Democratic Diffusion of AI", launched at the India AI Impact Summit and mentioned in the New Delhi Declaration, provides a foundation.[83] These multilateral efforts should be expanded to the agentic context, before the digital divide creates a rupture that becomes too large to fix.

7. Conclusion

Agentic systems have been the holy grail of computing since the 1950s. We are on the cusp of achieving it. The question today is not whether we can build AI systems that can think and act like humans, but whether our governance structures will crumble under the pressure of the coming technology revolution.

As this paper demonstrates, we are dealing with a new incarnation of AI. Agentic systems are highly capable, autonomous systems that can adapt to localised contexts, operate computers with human-like proficiency, and perform actions over long horizons in real-world settings, with potentially irreversible consequences.

The cascading risk framework proposed in this paper is not a theoretical construct. It is a description of what happens when we deploy powerful systems without adequate guardrails – failures that begin with one transaction, one compromised agent, one exposed database, spread outward in unanticipated ways. The five key governance issues identified in this paper — liability, oversight, access controls, cross-border coordination, and value alignment — are not unsolvable. However, addressing these issues will require concerted action, including new evaluation systems and benchmarks, common technical standards, risk mitigation methods and international coordination.

There is a small window to make this change. We need to build governance frameworks that account for the magnitude and intensity of change underway, while being mindful of the economic opportunities that the agentic era presents. We are not short of ideas. We are short of time.

8. References

1. Agentic systems can be understood as "systems that combine the intelligence of advanced AI models with access to tools so they can take actions on behalf of users, under their control" ("A Policy Framework for Agentic AI", Google, forthcoming publication)
2. Developing AI Agents: A Hands-On Guide with Real-World Applications, NASSCOM, March 2025, available at: https://community.nasscom.in/communities/ai/developing-ai-agents-hands-guide-real-world-applications
3. Agents, Robots and Us: Skill Partnerships in the Age of AI, McKinsey Global Institute, November 25, 2025, available at : https://www.mckinsey.com/mgi/our-research/agents-robots-and-us-skill-partnerships-in-the-age-of-ai#/
4. The author of this paper was the Lead Writer for the India AI Governance Guidelines (Ministry of Electronics and Information Technology, Government of India, "India AI Governance Guidelines: Enabling Safe and Trusted AI Innovation," November 2025, https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf) and part of the advisory group involving in planning the India AI Impact Summit, 2026.
5. The International AI Safety Report 2025 defines agents as "a general-purpose AI which can make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with its environment – for example by creating files, taking actions on the web, or delegating tasks to other agents – with little to no human oversight."; Anthropic defines agents as "systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks."; Vietnam's AI law defines agents as "highly autonomous AI systems capable of independently decomposing complex goals into sub-tasks and executing actions in digital or physical environments to achieve objectives without direct, continuous human intervention." See "International AI Safety Report 2025," January 29, 2025, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf; See "Building effective agents," Anthropic, December 19, 2024, https://www.anthropic.com/engineering/building-effective-agents; ; See Ministry of Science and Technology (Vietnam), *Law on Artificial Intelligence (Version 10)*, September 25, 2025, https://mic.mediacdn.vn/document/2025/10/2/250925duthao-luatai-v10-17593934466665893284209.pdf
6. Infocomm Media Development Authority (IMDA), "Model AI Governance Framework for Agentic AI," January 22, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf.
7. Kevin Feng, David McDonald, and Amy Zhang, "Levels of Autonomy for AI Agents," *Knight First Amendment Institute at Columbia University*, July 28, 2025, https://knightcolumbia.org/content/levels-of-autonomy-for-ai-agents-1.
8. Interview no. 4.
9. Interview no. 1.
10. Interview no. 7.

11. See Mark Stone, "2026 Microsoft Copilot Security Concerns Explained," *Concentric AI*, September 10, 2025, https://concentric.ai/too-much-access-microsoft-copilot-data-risks-explained/.
12. Interview no. 8, 9.
13. See Perplexity AI, "Introducing Perplexity Computer," *Perplexity Hub Blog*, February 25, 2026, https://www.perplexity.ai/hub/blog/introducing-perplexity-computer; See Alice Moore, "Perplexity Computer Review: What It Gets Right (and Wrong)," *Builder.io Blog*, March 4, 2026, https://www.builder.io/blog/perplexity-computer.
14. Interview no. 3.
15. Interview no. 11.
16. See Andrej Karpathy, "We're Summoning Ghosts, Not Building Animals," YouTube video, March 8, 2026, https://www.youtube.com/watch?v=lXUZvyajciY.
17. Interview no. 8.
18. Interview no. 11.
19. Interview no. 1.
20. Sanjeev Sanyal, Pranav Sharma, and Chirag Dudani, *A Complex Adaptive System Framework to Regulate Artificial Intelligence*, Economic Advisory Council to the Prime Minister Working Paper, January 2024, https://eacpm.gov.in/wp-content/uploads/2024/01/EACPM_AI_WP-1.pdf
21. Interview no. 9.
22. Interview no. 11.
23. Interview no. 5.
24. See Russell Brandom, "Zuckerberg Teases Agentic Commerce Tools and Major AI Rollout in 2026," *TechCrunch*, January 28, 2026, https://techcrunch.com/2026/01/28/zuckerberg-teases-agentic-commerce-tools-and-major-ai-rollout-in-2026/; See Mastercard, *Agentic Commerce, Q3 2025* (Mastercard Signals), https://view.ceros.com/mastercard-labs/mastercard-agentic-commerce-q3/p/4; See Kevin Ichhpurani, "Shaping the Future Together with Our Partners: The Potential of Agentic AI," *Google Cloud Blog*, July 17, 2025, https://cloud.google.com/blog/topics/partners/sharing-new-report-on-the-potential-of-agentic-ai.
25. NASSCOM, "*Enterprise Experiments with AI Agents – 2025 Global Trends*," Report, June 2025, https://community.nasscom.in/system/files/report/Enterprise%20Experiments%20with%20AI%20Agents%20-%202025%20Global%20Trends%20vF.pdf.
26. See Model Context Protocol, "Introduction to Model Context Protocol," *Model Context Protocol Documentation*, accessed March 8, 2026, https://modelcontextprotocol.io/docs/getting-started/intro; See A2A Protocol, "What is A2A?," March 8, 2026, https://a2a-protocol.org/latest/
27. For example through Salesforce Agentforce deployment, Reddit deflected 46% of support cases and cut resolution times by 84%, OpenTable resolved 70% of diner and restaurant inquiries autonomously, and Engine reduced handle time by 15%, saving over $2 million annually. See Salesforce, "Welcome to the Agentic Enterprise: With Agentforce 360, Salesforce Elevates Human Potential in the Age of AI," October 13, 2025, https://www.salesforce.com/ap/news/press-releases/2025/10/14/welcome-to-the-agentic-enterprise-with-agentforce-360-salesforce-elevates-human-potential-in-the-age-of-ai/ .

28. See John Werner, "How Agent Swarms Will Change the Web and Everything Else," Forbes, December 29, 2025, https://www.forbes.com/sites/johnwerner/2025/12/29/how-agent-swarms-will-change-the-web-and-everything-else/; See Matthew Sharp, Omer Bilgin, Iason Gabriel, and Lewis Hammond, "Agentic Inequality," arXiv, October 19, 2025, https://arxiv.org/html/2510.16853v2; See Muhammad Atta Ur Rahman and Melanie Schranz, "LLM-Powered Swarms: A New Frontier or a Conceptual Stretch?," arXiv, June 17, 2025, *https://arxiv.org/html/2506.14496v1*.
29. For example, users could use OpenClaw, an open-source personal AI agent framework that runs locally on your machine, enabling users to manage proactive AI assistants via chat apps like WhatsApp or Telegram. It supports custom tools (called "skills" or first-class agent tools) for actions like file management, browser control, shell execution, and integrations with services such as GitHub or Todoist. See "OpenClaw," *OpenClaw Documentation*, accessed March 8, 2026, https://docs.openclaw.ai/.
30. Jobs relating to data entry, accounting clerks, typists and word processing operators, statistical, finance, and insurance clerks, etc. are potentially at highest risk of automation with advancement in AI. See Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang, "Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce," *arXiv*, June 6, 2025, https://arxiv.org/pdf/2506.06576.pdf; See "How might generative AI impact different occupations?" *International Labour Organization*, May 20, 2025, https://www.ilo.org/resource/article/how-might-generative-ai-impact-different-occupations#occupations.
31. Jack Clark, "Import AI 443: Into the mist: Moltbook, agent ecologies, and the internet in transition," *Import AI,* Newsletter, February 2, 2026, https://importai.substack.com/p/import-ai-443-into-the-mist-moltbook.
32. Interview no. 6.
33. Empirical data supports the trajectory: an analysis of 900,000 newly created web pages found that nearly three-quarters contained AI-generated content. See Ryan Law, Xibeijia Guan, and Tim Soulo, "74% of New Webpages Include AI Content (Study of 900k Pages)," *Ahrefs Blog*, May 19, 2025, https://ahrefs.com/blog/what-percentage-of-new-content-is-ai-generated/.
34. Amlan Mohanty, "The Trust Inversion," *Techlawtopia*, Blog, October 30, 2025, https://www.techlawtopia.com/the-trust-inversion/.
35. "Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules", *Ministry of Electronics and Information Technology, Government of India*, notified February 10, 2026, https://www.meity.gov.in/static/uploads/2026/02/550681ab908f8afb135b0ad42816a1c9.pdf.
36. Instagram post by Adam Mosseri (@mosseri), December 31, 2025, https://www.instagram.com/p/DS7pz7-DuZG/.
37. In Sweden, government agencies, ministries, and municipalities use the standard country-code domain .se rather than a restricted, dedicated government TLD like .gov.in. In this context, agents may mistakenly trust a deceptive website as providing bonafide information from official sources.
38. Interview with Venkata

39. Anthropic, "Disrupting the first reported AI-orchestrated cyber espionage campaign," News Release, November 13, 2025, https://www.anthropic.com/news/disrupting-AI-espionage.
40. Ministry of Electronics and Information Technology, Government of India, "India AI Governance Guidelines: Enabling Safe and Trusted AI Innovation."
41. Reed Albergotti, "Anthropic's investors stay silent in fight against the Pentagon," *Semafor Technology*, March 4, 2026, https://www.semafor.com/newsletter/03/04/2026/anthropics-investors-stay-silent-in-fight-against-pentagon?utm_source=headernewsletterlink&utm_medium=technology.
42. See "Request for Information: Security Considerations for Artificial Intelligence Agents Docket No. NIST-2025-0035", Anthropic,March 9, 2026, https://www-cdn.anthropic.com/43ec7e770925deabc3f0bc1dbf0133769fd03812.pdf
43. Infocomm Media Development Authority (IMDA), "Model AI Governance Framework for Agentic AI."
44. "FREE-AI Committee Report - Framework for Responsible and Ethical Enablement of AI", Reserve Bank of India,August, 2025, https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/FREEAIR130820250A24FF2D4578453F824C72ED9F5D5851.PDF
45. Interview no. 5.
46. Interview no. 1.
47. Interview no. 3.
48. Interview no. 7.
49. Interview no. 9, 10, 5.
50. Interview no. 7.
51. Interview no. 5.
52. Interview no. 4.
53. Interview no. 8, 11.
54. Interview no. 6.
55. Interview no. 2.
56. See Usman Naseem, "Mechanistic Interpretability for Large Language Model Alignment: Progress, Challenges, and Future Directions," *arXiv*, February 12, 2026, https://arxiv.org/pdf/2602.11180; See Subramanyam Sahoo, Aman Chadha, Vinija Jain, and Divya Chaudhary, "Position: The Complexity of Perfect AI Alignment — Formalizing the RLHF Trilemma," *arXiv*, November 23, 2025, https://arxiv.org/pdf/2511.19504; See Anthropic, "Constitutional AI: Harmlessness from AI Feedback," *arXiv*, December 15, 2022, https://arxiv.org/pdf/2212.08073.
57. See Geetha Raju and Balaraman Ravindran, *AI Incident Reporting Framework for India*, Discussion Paper, Centre for Responsible AI, Wadhwani School of Data Science and AI and Indian Institute of Technology Madras, October 2025, https://cerai.iitm.ac.in/docs/AI_Incident_Reporting_V1.pdf

58. The EU AI Act classifies use-cases into unacceptable risk, high risk, limited risk, and minimal risk categories and prescribes varying levels of transparency, documentation and other obligations contingent on the use-case. Notably, use-cases in the unacceptable risk category, which include manipulative deception, biometric systems which infer demographic data, and untargeted scraping through facial recognition technology to create databases, are prohibited to deploy or bring to market. See Regulation (EU) 2024/1689 of the European Parliament and of the Council, Artificial Intelligence Act, 13 June 2024, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng .
59. See Yoshua Bengio et al., "*International AI Safety Report 2025*," January 29, 2025, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025.
60. See Anthropic, "Our framework for developing safe and trustworthy agents," *Anthropic*, News Release, August 4, 2025, https://www.anthropic.com/news/our-framework-for-developing-safe-and-trustworthy-agents; See Paula Goldman, "How Salesforce shapes ethical AI standards in the agent era," *Salesforce News*, October 28, 2024, https://www.salesforce.com/in/news/stories/responsible-agentic-ai-guidelines/; See Microsoft, "Apply responsible AI principles — Microsoft Copilot Studio guidance," *Microsoft Learn*, updated January 20, 2026, https://learn.microsoft.com/en-us/microsoft-copilot-studio/guidance/responsible-ai.
61. Interview no. 1, 5, 10.
62. Interview no. 8.
63. For instance, Anthropic and OpenAI presented different views on guardrails in the context of their negotiations with the U.S. Department of War in February 2026. See "Statement from Dario Amodei on our discussions with the Department of War," Anthropic, February 26, 2026, https://www.anthropic.com/news/statement-department-of-war/; See "Our Agreement with the Department of War," OpenAI, March 2, 2026, https://openai.com/index/our-agreement-with-the-department-of-war/.
64. Interview no. 3.
65. In comparison to previous voluntary commitments frameworks, the New Delhi Frontier AI Impact Commitments especially focus on collating information on the economic impact of AI systems, and on promoting multilingual AI offerings to meet varied socio-cultural contexts. See , Ministry of Electronics & IT, Government of India, "Championing Inclusive and Multilingual AI for the Global South, India Unveils New Delhi Frontier AI Commitments," Press Information Bureau, February 19, 2026, https://www.pib.gov.in/PressReleasePage.aspx?PRID=2230201&reg=3&lang=1.
66. See Model Context Protocol, "Introduction to Model Context Protocol" https://modelcontextprotocol.io/docs/getting-started/intro.; See A2A Protocol, "What is A2A?, https://a2a-protocol.org/latest/."
67. Interview no. 4.
68. Interview no. 5.
69. Interview no. 6.
70. See AI Agent Standards Initiative, Center for AI Standards and Innovation, https://www.nist.gov/caisi/ai-agent-standards-initiative.

71. See ISO/IEC 42001:2023, https://www.iso.org/standard/42001#:~:text=ISO/IEC%2042001%20is%20an%20international%20standard%20that,manage%20risks%20and%20opportunities%20associated%20with%20AI.
72. See Agentic AI Foundation website at https://aaif.io/
73. See Miles Brundage, et. al. "Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies", https://www.averi.org/ourwork/frontier-ai-auditing?utm_source=substack&utm_medium=email
74. Interview no. 2.
75. iSPIRT ProductNation, "DEPA AI Chain: Empowerment Through Provenance," January 10, 2025, https://pn.ispirt.in/depa-ai-chain-empowerment-through-provenance/.
76. See Office of the Principal Scientific Adviser to the Government of India, "Strengthening AI Governance Through Techno-Legal Framework," January 2026, https://psa.gov.in/CMS/web/sites/default/files/publication/AI-WP_TechnoLegal.pdf.
77. Interview no. 3.
78. See Paula Goldman, "How Salesforce shapes ethical AI standards in the agent era," *Salesforce News*, October 28, 2024, https://www.salesforce.com/in/news/stories/responsible-agentic-ai-guidelines/.
79. Interview no. 10.
80. "Techdirt Podcast Episode 427: Why Data Portability Is Crucial For The AI Future," Techdirt, August 19, 2025, https://www.techdirt.com/2025/08/19/techdirt-podcast-episode-427-why-data-portability-is-crucial-for-the-ai-future/.
81. See "Remarks by Director Michael Kratsios at the India AI Impact Summit," The White House, February 20, 2026, "As the Trump Administration has now said many times: We totally reject global governance of AI. We believe AI adoption cannot lead to a brighter future if it is subject to bureaucracies and centralized control." available at https://www.whitehouse.gov/articles/2026/02/remarks-by-director-michael-kratsios-at-the-india-ai-impact-summit/.
82. See "The Adolescence of Technology," Dario Amodei, January 2026, https://darioamodei.com/essay/the-adolescence-of-technology. Amodei argues that at current levels of development, AI systems will possess substantial scale and capabilities within the next few years, backed by necessary infrastructure, which could supercharge a country's economic and developmental trajectory. He describes this future as akin to a "country of geniuses in a data center" and argues that countries at the forefront of this trajectory are likely to enjoy concentrated benefits to the exclusion of others.
83. See Ministry of External Affairs, Government of India, "AI Impact Summit Declaration, New Delhi, February 18–19, 2026," https://www.mea.gov.in/bilateral-documents.htm?dtl/40809/.

About CeRAI

The **Centre for Responsible AI (CeRAI) at the Indian Institute of Technology, Madras**, is a multi-disciplinary, non-profit research centre positioned in the Global South, as one among the few global institutions that specialises in both technical and policy research to ensure and enable the responsible development and deployment of AI systems.