PARTNERSHIP ON AI

Preparing for AI Agent Governance

A research agenda for policymakers and researchers

Jacob Pratt
Thalia Khan

Contents

Key Takeaways **3**

Introduction **4**

What are AI agents? 4

Prioritizing policy-driven research 6

Policy questions for further research **7**

Understand the technology and policy landscape **7**

1.1 How are AI Agents different from existing AI systems? 8

1.2 Which existing legal and policy frameworks are relevant to AI agents? 9

1.3 How are different jurisdictions responding to AI agents? 9

Understand the risks and opportunities **10**

2.1 How are AI agents likely to be used and accessed? 10

2.2 How significant might benefits be from the adoption of AI agents? 11

2.3 How significant might risks be from the adoption of AI agents? 11

2.4 Where will the market naturally correct for any issues? 12

Understand the interventions **13**

3.1 How might policymakers implement innovation-enabling initiatives, such as regulatory sandboxes? 13

3.2 How might policymakers implement visibility, documentation, and post-deployment monitoring interventions? 14

3.3 How might policymakers incentivize or design infrastructure to support AI agent governance? 14

3.4 How might policymakers implement licensing, audit and assurance-related interventions? 15

3.5 How might policymakers leverage demand-side and supply-side levers to encourage beneficial AI adoption? 16

Monitoring, sandboxes, and testbeds as a policymaker priority **17**

PAI's focus on AI agents 17

Acknowledgements **19**

Endnotes **24**

Key Takeaways

- **How AI agents will impact society is still uncertain.** While the body of scholarly work and publicly available evidence[1] are growing, we don't yet know enough about how AI agents will be used or what impacts they may have. However, AI investments continue to soar, so policymakers should begin to prepare now.

- **Policymakers should prioritize evidence and information gathering, including through sandboxes and testbeds.** Given this uncertainty, policymakers should promote activities to generate evidence, rather than advancing prescriptive regulations. Promising options policymakers can use to build expertise and track developments are sandboxes and testbeds, which enable experimentation of new systems under regulatory supervision.

- **Subsequent rule-making will require substantial research from inside and outside of government.** Academia, civil society, government, and industry should work together to generate evidence on AI agents' capabilities, risks, societal impacts, and potential policy interventions. This will support future policy development.

- **This paper provides a roadmap for this research.** We outline three foundational requirements for governing AI agents and detail a comprehensive research agenda, including 12 top-level and 45 sub-level questions, designed to directly support policymakers in developing evidence-based policy.

Introduction

As governments grapple with governing generative AI systems, AI agents — systems built on foundation models that can autonomously complete complex tasks by developing plans and accessing tools to take actions digitally — are being seen as the next frontier for AI systems. Industry leaders have named 2025 as the "year of agentic exploration," foretelling the adoption of systems that will change how we interact, what jobs we perform, and even how we think. Some of these impacts are not just theories, but are being felt now.[B]

B For example, decreases in software developer roles are partially being explained by increased AI agent usage (Computer World, 2025)

Despite these high expectations, widespread AI agent adoption has been stifled by persistent reliability and security challenges. Although Gartner predicts that 40% of agentic AI projects will fail by 2027, AI funding has continued to boom. Recent international policy proposals have focused on promoting the adoption of AI to compete on the international stage, suggesting that advancements in AI agent engineering will lead to significant real-world impact.

Policymakers now have an opportunity to support the safe and responsible scaling of AI agents while preparing for their impacts. Unlike with generative AI systems, which have transformed our society faster than governance frameworks can be adapted, policymakers have an opportunity to prepare thoughtfully for how to promote the benefits and protect against the risks of AI agents. The key challenge for policymakers is not whether to regulate now, but how to govern AI agents while their impacts are still uncertain, and understand what evidence will be needed to make informed decisions when decisive action is needed.

The key challenge for policymakers is not whether to regulate now, but how to govern AI agents while their impacts are still uncertain.

What are AI agents?

While generative AI systems produce content for humans to act on, agents — built on the same foundation models with added scaffolding — reason, plan, and perform sequences of actions to achieve user goals. Unlike generative AI, these systems directly execute actions by using digital tools to interact with complex environments. We are already seeing prototypes of agents that can schedule meetings through a calendar API or book flights via web interfaces. More ambitious proposals include agents that negotiate contracts, assist in healthcare decisions, and coordinate supply chains.

For the purposes of the agenda we will focus on systems that execute direct actions with minimal human oversight, as defined by Levels 3-5 of environmental interaction outlined by Srikumar, et al.[2] (see Figure 1).[C] As these systems advance, "constrained" agents will pave the way for more adaptive unconstrained agents, which can have greater potential impacts.

C We also encourage you to explore the Ada Lovelace Institute's Delegation Nation, IBM's What are AI Agents?, and CDT's AI Agents in Focus, Technical and Policy Considerations to understand the basics of AI agents.

Figure 1. Levels of agent influence on digital environments, with examples of LLM-based systems.
Current systems operate at Levels 1–3, while Levels 4–5 illustrate emerging directions for more autonomous agents.

	LEVEL	ABILITY TO INFLUENCE THE ENVIRONMENT	EXAMPLE AI SYSTEMS
	0	Read-only, observation only	Speech recognition systems Image classification models
MEDIATED INFLUENCE	1	**Mediated influence via humans:** System outputs text suggestions or advice; output only affects the environment if a human acts	ChatGPT without tools
MEDIATED INFLUENCE	2	**Mediated influence via humans with passive tools:** System uses tools like search or knowledge databases for context, not for changing the world itself	ChatGPT or Claude with web search *(provides info but doesn't act)* Gemini Deep Research *(provides research analysis)* GitHub Copilot *(provides code suggestions)*
DIRECT ACTIONS	3	**Direct Actions — Constrained Agent:** System performs single-step actions directly using predefined tools, acting on user commands without needing humans to carry out the result	Operator Claude Computer Use Manus Project Mariner Cursor AI Other code-executing or API-calling agents
DIRECT ACTIONS	4	**Direct Actions — Semi-constrained Agent:** System accepts broad goals, decomposes them into multi-step plans, and executes steps autonomously across known tools, *without needing humans to approve every step.*	Longer and more complex workflows. E.g. User says "File my taxes" → the agent gathers necessary documents from user's email, fills tax forms, and submits them using chosen e-filing service, without asking for approvals on each step.
DIRECT ACTIONS	5	**Direct Actions — Unconstrained Agent:** System accepts a broad goal, autonomously executes multi-step plans, and adapts by integrating new tools or strategies beyond what a user configured, all without requiring approvals.	Adaptive workflows. E.g. User says "File taxes for my business" → the agent autonomously gathers financial records, contacts suppliers for missing information, interprets regulations, applies business-specific deductions, switches between tools or services as needed, and completes the filing with without further human input.

We assess levels of capability based on environmental interaction, but an agent's ability to interact with its environment also depends on other differentiating properties: whether it relies on predefined tools or can flexibly adapt tool use, how much human oversight it requires, and the complexity of goals it can pursue. These properties vary by degree and collectively shape an agent's ability to interact with its environment. Our levels share similarities with recent surveys of agent autonomy, which highlight dimensions such as constraints on environmental impact and flexibility of actions.[3] Although some systems might meet our Level 3 criteria through rule-based execution (e.g., spam bots), our analysis focuses on LLM-driven agents using reasoning, planning, or decision-making, as these introduce new sources of unpredictability, runtime failure, or hazards. Levels 4 and 5 draw inspiration from Patel, 2025.[4]

Prioritizing policy-driven research

These increasingly unconstrained agents present new challenges for policymakers. Unlike generative AI systems, which generate an output but require people to take further actions, agents can directly interact with the environment.[D] These systems can help companies optimize supply chain management and execute financial trades, but can also lead to the system impersonating users or making unauthorized transactions.

The growing deployment of these AI agents will intersect with the core remits of many government departments and regulators and create profound opportunities and risks. The U.S. Department of Labor, tasked with advancing "opportunities for profitable employment," must contend with the potential for agents to both displace workers through automation and create new opportunities through upskilling. Similarly, the UK's Information Commissioner's Office, which guards against "high risk to individuals and their rights," must prepare for novel privacy violations that could occur through agent failures. Departments, regulators, and other policymakers will need to take action to progress their missions in the face of this technological advance.

Policymakers will require substantial research and evidence to make the right public policy decisions on AI governance — including on the institutions, policies, regulations, and tools that ensure that AI systems operate in the public interest.[E] However, advocating for significant investment to govern AI agents before widespread adoption occurs and harms are realized is politically challenging, especially when this investment can address harms from existing technologies.[5]

The scientific community across academia, civil society, and industry can help by conducting research on AI agents. To coordinate this effort, we outline a research agenda designed explicitly to support public policymaking for AI agents. Developed through a comprehensive literature review and a workshop with over 30 cross-sector experts, this agenda focuses on actionable lines of research that can inform the development of effective policy, governance, and regulation.

NOTE
Though this research agenda is designed to support policymaking, we do not expect policymakers to conduct research relevant to all the questions listed. Prioritizing government-led research on AI agents is vital for progressing public interest research, but will mean deprioritizing other areas, which might not be feasible. This Research Agenda encourages private organizations to conduct research that furthers their own goals while also contributing to public governance. However, it may be beneficial for governments to identify key questions based on their own priorities that are worth exploring. We describe how sandboxes and testbeds may be a useful first step for policymakers later in the report.

D Though previous systems have also taken automated actions to affect the environment, such as automated financial trading systems, these systems have acted within constrained boundaries to complete narrow tasks. AI agents build on foundation models to complete a variety of complex tasks, leading to new risks and opportunities.

This agenda focuses on actionable lines of research that can inform the development of effective policy, governance, and regulation.

E There have been calls for evidence-based AI policy (Bommasani R. et al., A Path for Science- and Evidence-based AI Policy, 2024) and a recognition that this should not be used to downplay urgency and delay policymaking (Caspar S. et al, Pitfalls of Evidence-Based AI Policy, February 2025). We lay out our argument for needing additional research in this paper.

Policy questions for further research

To identify what evidence is needed to support policymakers, we conducted a comprehensive literature review and convened a workshop with over 30 experts from academia, civil society, industry, and government. This resulted in an agenda describing three core requirements for policymakers to govern AI agents effectively, with 12 top-level questions providing more detail on lines of research.

There is significant overlap between the requirements and top-level questions for AI agents and existing AI systems, as AI agents are a "normal technology" and not a radical new force. This means policymakers should not take a radically different approach to governing them. But the details of policy implementation will be different, so we also identify 45 sub-level questions and provide context on how these questions describe challenges for policymakers unique to AI agents, including through the hypothetical use case of an "unconstrained" level 5 tax-filing agent.[F]

This framework provides direction for future research by academic, civil society, government, and industry actors. PAI will be progressing this research agenda, focusing on key risks to human connection, labor and the economy, and risks from agent failures, as well as key policy interventions related to international governance, and sandboxes and testbeds.

F This agent can:
- Access your financial accounts.
- Aggregate your financial information.
- Research and interpret tax-filing guidance.
- Develop a plan for completing your tax return.
- Create tax documents and complete tax forms.
- Audit your submission.
- Submit documents to the relevant authorities.

This hypothetical example draws inspiration from Dwarkesh Patel, Why I Don't Think AGI Is Right Around the Corner.

REQUIREMENT 1

Understand the technology and policy landscape

> "How do (or perhaps, how should) our existing laws, regulations, and institutions that govern individuals and organisations apply to the agents acting on their behalf? (How) should we modify, augment, or replace them?"
>
> —EXPERT PARTICIPANT, PAI POLICY WORKSHOP, JULY 2025

Policymakers need a clear model for thinking about AI systems when designing frameworks.[G] A common language to discuss components, safety mechanisms, and actors in the value chain[H] would enable collaboration between stakeholders.

This will contribute to developing a solid understanding of how AI agents function and interact within existing regulatory frameworks, without which policymakers risk duplicating or developing conflicting rules.[I]

G See NIST's AI RMF, Figure 3 and characteristics of trustworthy AI for examples.

H See PAI's Risk Mitigation Strategies for the Open Foundation Model Value Chain for an example value chain.

I The impacts of conflicting rules are discussed in PAI's Policy Alignment on AI Transparency report.

1.1 How are AI Agents different from existing AI systems?[6]

AI agents differ from existing AI systems, with Kasirzadeh A. & Gabriel I. noting four noticeable differentiating characteristics:

- **Autonomy:** The ability to perform actions without external direction or control.
- **Influence on the environment:** The ability to perceive and causally impact the environment.[J]
- **Goal complexity:** The ability to develop and execute plans to achieve complex goals.
- **Generality:** The ability to operate successfully across different tasks.

Understanding these differences is crucial for policymakers because systems with greater capabilities, especially autonomous ones, carry higher risks. By understanding how these systems differ, policymakers can create more effective regulations to manage the evolving risks associated with AI agents.

J "The environment" relates to the parts of the world that an agent can perceive and act upon. For a generative AI system, this environment may be limited to chat interactions with users, though indirect environmental effects can occur through humans. However, AI agents can shape the environment more directly via computer operations, such as application programming interface (API) calls.

SUB-QUESTIONS

1.1.1 What are the unique characteristics of AI agents that generate the need for policymaking?

1.1.2 How do the levels of characteristics of an AI agent affect the level or type of governance required? What characteristic thresholds are required before governance measures apply?

1.1.3 What is the correct level of abstraction to 'model' the key components of AI agents? How should relevant actors be defined?

HYPOTHETICAL USE CASE

A tax-filing agent demonstrates key differences from current AI systems, which will have policymaking implications. While a generative AI system might simply answer tax questions, the agent might autonomously navigate multiple financial accounts, interpret complex IRS regulations, and execute multi-step workflows without constant oversight. Unlike traditional rule-based tax software, it might reason through edge cases and coordinate across multiple tools and data sources to complete the entire filing process end-to-end.

1.2 Which existing legal and policy frameworks are relevant to AI agents?[7]

"New" technologies, like AI agents, are rarely completely unique, and their development and deployment will typically fall under the scope of existing regulations or policies.[K] However, the increased autonomy of AI agents has implications on the efficacy of existing legal frameworks, particularly in tort and agency law. It is important that policymakers understand how existing laws apply to autonomous systems to avoid creating redundant or conflicting legislation; making more informed and targeted legislative decisions.

K For example, the EU AI Act recognizes that providers of certain AI systems also have to adhere to other EU legislation, including related to data protection (Regulation (EU) 2016/679), financial services (Directive 2013/36/EU), and medical devices (Regulation (EU) 2017/745).

SUB-QUESTIONS

How does existing policy apply to AI agents, including:

- **1.2.1** General/non-AI specific legal frameworks, such as tort law and product liability?
- **1.2.2** Specific AI-related policy frameworks, such as the EU AI Act or NIST's AI Risk Management Framework?
- **1.2.3** Sector-specific regulations, such as financial regulation?
- **1.2.4** What policy and legal instruments will be needed (or adapted) in a world with mass deployment of advanced agent systems?
- **1.2.5** Which laws or regulations might be suitable for aligning AI agents to? How might this be done?

HYPOTHETICAL USE CASE

Existing legal and policy frameworks may apply to tax-filing agents, which will require research to clarify for policymakers. Liability questions might arise when the agent makes errors, with the user, developer, or other actors potentially responsible for penalties. Accessing financial data may trigger privacy regulations, and any AI-powered advice may fall under consumer protection and financial services regulations.

1.3 How are different jurisdictions responding to AI agents?[8]

PAI's analyses "Policy Alignment on AI Transparency" and "Decoding AI Governance" describe how current AI policies overlap and support each other; a similar approach is needed for policies concerning AI agents. By understanding the existing regulatory landscape, policymakers can avoid duplicating efforts and creating conflicting regulations or legislation, and support greater interoperability and international cooperation which will benefit people on a global scale.

SUB-QUESTIONS

1.3.1 What AI agent specific policies and legislation are being developed in other jurisdictions? What guidance is being provided on how to apply existing policies and legislation?

1.3.2 How can the AI Governance stack encourage interoperability across jurisdictions?

HYPOTHETICAL USE CASE

Policymakers should look to coordinate internationally to prevent regulatory fragmentation that could create compliance burdens for developers. For example, the EU may classify a tax-filing agent as high-risk, while the US may rely on existing IRS regulations, and so joint tax-filing across both jurisdictions may face issues.

REQUIREMENT 2

Understand the risks and opportunities

> "Most unreliability issues will be solved by market incentives—if they are visible. Which might not be visible and require regulator intervention?"
>
> —EXPERT PARTICIPANT, PAI POLICY WORKSHOP, JULY 2025

Policymakers will need to take action when the impact of a technology affects their policy goals. This requires an understanding of what these impacts are — the potential scale of benefit from the adoption of AI agents, and the severity and likelihood of risks related to these systems. This is determined by where AI agents are likely to be deployed and used, and if AI agents do not impact a government organization's objectives, then policymakers may wish to focus on other priorities.

Detailing all the theoretical risks and benefits associated with AI agents is out of scope for this paper.[L] Instead, we highlight key risks and opportunities mentioned through multistakeholder engagement (Q 2.2.2 - 2.2.4 and Q 2.3.1 - 2.3.6) and supporting questions identified as priorities by workshop participants (such as Q 2.3.7 and Q 2.4).

L The Ethics of Advanced AI Assistants, International AI Safety Report and MIT's Domain Taxonomy of AI Risks describe taxonomies of risks and benefits associated with AI and AI agents.

2.1 How are AI agents likely to be used and accessed?[9]

While AI agents have the potential to be adopted for a variety of tasks, we have yet to see widespread adoption (as of September 2025). Anticipating how AI agents will be used requires examining research, current deployment patterns, and emerging use cases. As PAI's Documenting the Impacts of Foundation Models report details, this can support policymakers with planning how to respond.[M] Identifying the sectors where AI agents are being adopted most rapidly can also help policymakers plan a response, create targeted policies, and fund relevant research or guidance.

M For example, the UK's Bank of England shared a report detailing that AI may begin to impact core financial decisions, and detailed how they plan to monitor and respond to these emerging uses.

SUB-QUESTIONS

2.1.1 For which tasks is there likely to be wider usage of AI agents? How might this vary by domain, sector, use case, demographic group, and region?

2.1.2 Where is there likely to be a "digital divide" which blocks beneficial usage?

2.1.3 How can policymakers anticipate and influence future impacts from AI agents, such as using foresight techniques?

HYPOTHETICAL USE CASE

Early adoption of tax-filing agents may focus on simple tax returns before expanding to more complex scenarios, and usage patterns may spike during tax season, requiring scalable infrastructure. Policymakers should promote equitable access to these agents to prevent a digital divide where only wealthy taxpayers benefit, and so could improve access by piloting low-income taxpayer assistance programs.

2.2 How significant might benefits be from the adoption of AI agents?[10]

It's theorized that AI agents could provide major societal benefits, such as supporting medical R&D and enhancing fraud detection. Understanding the scale of these benefits, where these benefits may be realized, and by whom can help policymakers support and amplify their realization, such as through targeted education initiatives or research funding. With effective governance and assurance, the increased autonomy, efficacy, generality, and ability to complete goals related to more complex tasks will enable these systems to realize benefits that support a just, equitable, prosperous society.[N]

N Characteristics are detailed in Characterizing AI Agents for Alignment and Governance (Kasirzadeh A. & Gabriel I., 2025.)

SUB-QUESTIONS

2.2.1 How can we measure the benefits from adopting AI agents? How is this distributed between organizations and people, including SMEs, government agencies, and the general public?

2.2.2 For which domains, sectors, or use cases is there likely to be significant benefits from adopting AI agents?

How significant might benefits be in areas of public interest, such as, but not limited to:

2.2.3 Public sector usage?

2.2.4 Healthcare?

2.2.5 Education?

2.2.6 What might prevent people from benefitting from AI agents?

HYPOTHETICAL USE CASE

Policymakers should look to maximize the benefits of tax-filing agents. For example, tax-filing agents could improve access to tax-filing support, maximize legitimate tax deductions, and reduce tax preparation costs for citizens.

2.3 How significant might risks be from the adoption of AI agents?[11]

A vast array of literature details risks of AI agents,[o] from multi-agent interactions and alignment and misuse to risks from specific domains, such as cybersecurity and finance. We are also beginning to see case studies of actual harms and industry responses, and are expecting to see regular updates on the state of research detailing the risks from AI, including AI agents, as part of the International AI Safety Report series.

This literature shows how the increased ability for these systems to influence the environment may cause large scale negative impacts. Policymakers have duties to protect citizens from these harms, and anticipating the likelihood and severity of risks from AI agents is crucial to upholding this duty.

[o] We do not list out the entire list of risks and opportunities from the adoption of AI agents due to length (MIT's Domain Taxonomy of AI Risks lists 24 subdomains and identified over 1600 risks in literature). Instead, we aim to highlight high-level groupings of risks that are exacerbated by, or unique to, the novel capabilities of AI agents, or risks that PAI has identified as important and are actively exploring.

SUB-QUESTIONS

How significant might risks be from:

- **2.3.1** The interactions between multiple agents?
- **2.3.2** Agent failures, malfunctions, and accidents?
- **2.3.3** Agent misuse and malicious intent?

How significant might risks be to:

- **2.3.4** The information and trust ecosystem?
- **2.3.5** Labor and the economy?
- **2.3.6** Individuals and communities?

2.3.7 What are the pathways to causing harm? (i.e. threat models)

HYPOTHETICAL USE CASE

Policymakers may need to respond to some of the risks posed by tax-filing agents. For example, tax-filing agents may file returns incorrectly, leading to penalties, may use unauthorized tax strategies that trigger audits, and may expose sensitive financial information through prompt injection attacks. A systemic risk may include common failures that affect thousands of returns.

2.4 Where will the market naturally correct for any issues?[12]

It will be beneficial for industry organizations to put R&D efforts into mitigating some risks, such as reliability issues, as it will improve the product and provide a competitive advantage. However, policymakers will need to take action where the market fails, or implement regimes that incentivize industry-led governance, such as assurance and certification. Further research is required to understand exactly what the market will naturally correct for, and where gaps will be left and policymakers should intervene.

SUB-QUESTIONS

2.4.1 What AI agent infrastructure will be provided privately, and where will government intervention be needed?

2.4.2 Which risks, such as reliability, will be mitigated by the market, and which risks will require government intervention?

HYPOTHETICAL USE CASE

The market may address accuracy and reliability issues through competition, as users abandon error-prone services. However, the market may not naturally address data privacy standards, or ensure equitable access for low-income users. Addressing these market failures could be part of the role policymakers play.

REQUIREMENT 3

Understand the interventions

> "How can we get better transparency and disclosures from developers? Without better understanding of what's 'under the hood,' good policy design seems difficult."
>
> —EXPERT PARTICIPANT, PAI POLICY WORKSHOP, JULY 2025

Understanding the "toolbox" of policy interventions is crucial to ensuring that any policy-maker option assessment is thorough and evidence-based. Many actors are researching technical solutions to AI safety and risk mitigation questions[13] and these should be supported by research on the appropriateness of sociotechnical and regulatory initiatives.

We note that this is not a comprehensive list of interventions, and we do not cover sovereign AI, data & IP, security, healthcare, basic science funding, export controls, and other topics in detail.[P] This allows us to share sub-questions on the topics we do cover, enabling a greater level of precision for future work.

P IFP's analysis of comments on the US AI Action Plan highlights that recommendations from submissions covered the topics listed above.

3.1 How might policymakers implement innovation-enabling initiatives, such as regulatory sandboxes?[14]

Given the uncertainty around AI agent capabilities and current adoption-focused AI policy frameworks, policymakers should be looking to better understand the capabilities of AI agents and encourage innovation. Government organizations can run sandboxes, testbeds, or live testing initiatives to test AI agents in real-world or simulated conditions. This approach can enable the government to better understand the technology and viability of a planned or implemented policy, while also enabling industry organizations to accelerate their access to the market.

SUB-QUESTIONS

3.3.1 How can policymakers use regulatory sandboxes, testbeds, and live testing to assure systems and test the feasibility of policies?

3.3.2 How can policymakers create simulated environments for AI agents to test capabilities in high-stakes or complex regulatory environments?

HYPOTHETICAL USE CASE

Policymakers could use controlled testing environments to understand revenue impacts and compliance risks before widespread deployment. For example, a government department, such as the IRS, could establish a sandbox allowing limited deployment of tax-filing agents for specific taxpayer segments with close monitoring. Participants could submit test returns parallel to traditional filing, comparing accuracy and compliance. The sandbox could test agent interactions with current government systems and evaluate government audit triggers.

3.2 How might policymakers implement visibility, documentation, and post-deployment monitoring interventions?[15]

Addressing the risks of AI agents requires visibility: information about where, why, how, and by whom AI agents are used."[16] This information — which can be gathered through documentation, post-deployment system monitoring and logging, and incident reporting — helps evaluate existing governance structures, adapt these structures, and ensure the accountability of key stakeholders. However, this information can highlight harms caused by an agent, which industry organizations may want to avoid sharing information on, so policymakers may need to play a role in incentivizing and designing these systems.

SUB-QUESTIONS

3.2.1 How can policymakers ensure that design attributes that are important for governance are sufficiently documented and made transparent?

3.2.2 What metrics will support the monitoring of society wide, multi-agent risks, and how can policymakers monitor these metrics through networks in a privacy-preserving way? (For example the frequency and proportion of human-agent and agent-agent interactions)

3.2.3 How might policymakers incentivize and design post-deployment monitoring and incident reporting?

HYPOTHETICAL USE CASE

Policymakers should ensure accountability when AI errors lead to taxpayer penalties or missed revenue. Tax-filing agents could require comprehensive logging of decision rationales, data sources, and tools accessed. Real-time monitoring could flag unusual behavioral patterns before submission. Post-deployment tracking could identify systemic errors across returns, enabling fast corrections, and incident reporting could capture and communicate agent failures. Documentation requirements might include audit trails showing how the agent interpreted specific tax situations, supporting accountability when disputes arise.

3.3 How might policymakers incentivize or design infrastructure to support AI agent governance?[17]

It is beneficial for industry organizations to build infrastructure that enables agents to take more actions and be used by more people. However, there is also the need for infrastructure that enables governance by non-industry actors, such as methods for agent attribution, and mediates how agents interact with their environment, such as rollbacks. These may require policymaker support.

SUB-QUESTIONS

3.3.1 How can policy actors build and incentivize the development of AI agent trust infrastructure? For example, supporting infrastructure to enable agent IDs.

3.3.2 How can policy actors incentivize and implement agent "attribution" infrastructure that attributes actions to agents, mediates interactions, and detects and remedies harmful actions?

3.3.3 How can policy actors incentivize and implement agent "oversight" infrastructure that generates useful information and supports accountability?

3.3.4 How can this infrastructure build on existing cybersecurity and information technology infrastructure?

3.3.5 As the development of "agent channels" may be time-sensitive, what policies should govern the use of these channels?

3.3.6 How can international cooperation facilitate the development of global standards and infrastructure for AI agents, learning from organizations like IETF?

3.3.7 Which institution is best placed to lead technical and/or policy interoperability for agent governance?

HYPOTHETICAL USE CASE

Policymakers should aim to prevent fragmentation and ensure secure, coordinated interactions, especially with government systems. Standardized APIs could enable secure tax-filing agent interactions with government and financial databases and systems. Authentication protocols could verify agent identity and authorization levels. A centralized registry could track certified agents, their capabilities, and compliance history. "Circuit breakers" may prevent mass submissions of incorrect returns, rollback mechanisms may correct for widespread errors, and secure channels for agent-to-agent coordination could support joint returns.

3.4 How might policymakers implement licensing, audit and assurance-related interventions?[18]

In the same way that cars, legal professionals, and financial organizations require a license to operate, autonomous AI agents could require licenses, registration, or other assurances before being deployed. These assurances will need to be founded on agreed science and metrology standards, which are not yet in place and may require policy stimulus. Additionally, they should be focused on measuring real-world impacts from specific tasks.

SUB-QUESTIONS

3.4.1 How can policy actors encourage the development and usage of evaluations and benchmarks for AI agents that more directly correlate with real-world impact, compared to current practices?

3.4.2 How can policymakers create simulated environments for AI agents to test capabilities in high-stakes or complex regulatory environments?

HYPOTHETICAL USE CASE

Policymakers should protect taxpayers from unreliable or malicious agents, similar to existing protections for human tax preparers. Tax-filing agents could require certification similar to other tax professionals. Licensing tiers might distinguish basic agents (simple returns) from advanced agents (complex business filings). Each agent could have a specific agent ID to track certifications. Audit requirements could include third-party assessments of accuracy rates and security practices.

3.5 How might policymakers leverage demand-side and supply-side levers to encourage beneficial AI adoption?[19]

Policymakers across the globe are not only looking to reduce the risks of AI, but increase the diffusion and adoption of AI.[Q] Policymakers can create market signals to encourage people to use AI agents, such as by educating critical infrastructure employees on where agents may be useful, and fostering the wider industry that makes AI possible, such as by providing healthcare datasets that enable foundation models to better direct agent actions in that sector. These actions will support the adoption of AI agent systems.

Q Examples include the EU's AI Continent Action Plan, Japan's AI Promotion Act and the Brazilian AI Programme.

SUB-QUESTIONS

3.5.1 How might governments encourage beneficial AI adoption through demand-side policies, such as encouraging standards and interoperability and investing in AI education initiatives?

3.5.2 How might governments encourage beneficial AI adoption through supply-side policies, such as supporting access to data?

3.5.3 How might existing mechanisms, such as labor organizations, professional standards, and tax structures support governments in 3.5.1 and 3.5.2?

HYPOTHETICAL USE CASE

Policymakers should aim to accelerate beneficial adoption and ensure the benefits of agentic tax assistance are distributed widely. Demand-side policies could include tax credits for using certified agents or free agent access for low-income taxpayers. Supply-side support might involve grants for developing accessible interfaces, funding for tax law training datasets, or partnerships with community organizations. Government departments could provide official guidance on agent requirements, reducing development uncertainty.

The need for coordination

We call attention to the need for coordination when researching many of these questions, and encourage further research to identify specific U.S. and international agencies that should be tasked with interpreting any research, as a cross-government approach may be needed. For example, the agency tasked with exploring the need for agent ID or licensing interventions might be different from the agency with the mandate to support the infrastructure that enables this, which might be different from the agency who supports its standardization. Government coordination will be necessary, both nationally and internationally.[R]

R The importance of interoperability is noted in PAI's Policy Alignment on AI Transparency. However, Kerry C. et al. describe how we should not aim for the consolidated global governance of AI systems, and that networked efforts may be more desirable.

Monitoring, sandboxes, and testbeds as a policymaker priority

Developing this research agenda has made clear that many of these questions are interrelated, so policymakers may find that identifying a clear sequence of priorities and actions is challenging. However, combatting the current uncertainty around AI agents should be a priority, and policymakers must develop the capability to monitor these systems and their impacts in a way that complements current adoption-focused policy frameworks. Governments should invest in institutional capacity and AI talent to make this happen, but this may require significant funding that is not politically feasible to commit, so alternative approaches should be explored.

Sandboxes and testbeds — controlled environments for testing, trialing, and experimentation under regulatory supervision — provide a modern, pragmatic solution for policymakers, and have already been explored in the UK, Singapore, and other regions. Developing this approach should be the first step for policymakers, as they can enable policymakers and researchers to gather information on risks, benefits, and other interventions.

Running sandboxes and testbeds with industry and non-industry actors will still have an associated cost, but they will build institutional capacity at a lower cost than large upfront investments. This approach will ensure guardrails and governance work in the real-world and expand on specific sections of the US's AI Action Plan, the US's proposed SANDBOX Act, and EU's AI Act. This approach will also ensure that future policymaking is built on a deep understanding of these systems and practical experience. It will also evolve with AI agents but must be designed to combat regulatory capture.

PAI's focus on AI agents

Ensuring that sandboxes and testbeds help policymakers better understand agents and potential policy interventions, encourage innovation from industry, and support academic research will require involvement from across the ecosystem. We believe that research with our multistakeholder community is key to making this work, and that we can learn lessons from previous research[S] and regulatory initiatives[T] to deliver actionable guidance for policymakers on how to make these fit for AI agents.

Beyond regulatory sandboxes and testbeds, we will also build on our previous work in the fields of policy, media integrity, labor and the economy, and safety-critical AI to explore the following areas:

S For example, Ranchordas S. and Vinci V., 2024 and Jeník I. and Duff S., 2020.

T For example, GOV.UK's New Smarter Regulatory Sandbox developed to increase compliance, and Singapore's three sandboxes on GenAI, AI Assurance, and Privacy Enhancing Technologies.

1. **International governance of AI agents** SEE Q1.3

As AI agents increase in popularity around the world, risks that cross national borders or concern the international community as a whole — such as critical infrastructure disruption, privacy breaches, and electoral interference — can be exacerbated. PAI is exploring how to manage those risks leveraging existing global governance tools, such as international law, non-binding global norms, and international accountability mechanisms, and what changes might be needed to strengthen them as AI agents become more widespread. You can read more in our recent work on AI Agents & Global Governance: Analyzing Foundational Legal, Policy, and Accountability Tools.

2. **AI agents and human connection** SEE Q2.2: BENEFITS / Q2.3: RISKS

Increasingly personlike AI agents are more persuasive, emotionally evocative, and interactive, making them seem genuinely "personlike" to users. PAI is exploring how AI can support informed and connected communities by cultivating a field of key players from media, academia, civil society, industry, and philanthropy to align on norms for knowledge- and connection-affirming AI, offering a path forward for how to drive towards AI and connection benchmarks in practice.

3. **AI agent failures and monitoring** SEE Q2.3: RISKS / Q3.2: POST-DEPLOYMENT MONITORING

As agents take direct actions on the environment, they create new risks and require real-time failure detection. We have explored why and when real-time failure detection matters in our recent work on Prioritizing Real-Time Failure Detection in AI Agents, and will be continuing to explore how this can be operationalized.

4. **AI agents and labor** SEE Q2.2: BENEFITS / Q2.3: RISKS

We will also be exploring the impacts of AI agents on workers and the economy, building on our previous work in the area.

We look forward to exploring these questions with our partnership, policymakers, and the wider community. If you would like to be involved in our policy work on AI agents, please email jacob@partnershiponai.org.

Acknowledgements

This report was prepared with guidance from PAI's Policy Steering Committee (members who attended additional discussions are indicated by an asterisk). We appreciate the invaluable input provided by experts who participated in workshop discussions, including:

Aaron Martin
University of Virginia

Ahmed Saleh
Member of the African Union Advisory Group on AI

Angela Kane
Vienna Center for Disarmament and Non-Proliferation

Arianna Manzini
Google Deepmind

Bill Thompson
BBC

Christina J. Colclough
The Why Not Lab

Dan Treacher
Financial Conduct Authority

David Wakeling*
A&O Shearman

Deon Woods Bell*
Gates Foundation

Edmund Towers
Financial Conduct Authority

Elham Tabassi*
The Brookings Institution

Harry Farmer
Ada Lovelace Institute

Ian Eisenberg
Credo AI

Jam Kraprayoon
Institute for AI Policy and Strategy

Jamie Bernardi
Centre for the Governance of AI

Joelle Pineau
McGill School of Computer Science

Lama Nachman
IBM

Laurence Diver
Financial Conduct Authority

Leo Peyronnin
Omidyar Network

Lewis Hammond
Cooperative AI Foundation (CAIF)

Lisa Titus
Meta

Merlin Stein
UK AISI

Nicol Turner Lee
The Brookings Institution

Noam Kolt
Hebrew University

Peter Cihon
Johns Hopkins University

Peter Slattery
MIT FutureTech

Ruchika Joshi
CDT

Rumman Chowdhury*
Humane Intelligence

Sebastian Hallensleben*
Resaro

Sebastien Krier
Google Deepmind

Shameek Kundu
Infocomm Media Development Authority

Siddharth de Souza
Warwick University

Vasilios Mavroudis
The Alan Turing Institute

William Bartholomew
Microsoft

Special thanks to PAI staff who have been integral to the development of this report:

Claire Leibowicz, John Howell, Madhulika Srikumar, Neil Uhl, Rebecca Finlay, Stephanie Bell, Stephanie Ifayemi, Talita Dias

Endnotes

1 Anthropic, "System Card: Claude Opus 4 & Claude Sonnet 4", May 2025; Anthropic, "Economic Index", Accessed September 24 2025; Chatterji, Aaron et al., How People Use ChatGPT, September 2025; Mitchell, Margaret et al., Fully Autonomous AI Agents Should Not be Developed, February 2025; Chan, Alan et al., Harms from Increasingly Agentic Algorithmic Systems, May 2023.

2 Srikumar, Madhulika et al., Prioritizing Real-Time Failure Detection in AI Agents, Partnership on AI, September 2025.

3 Cihon, Peter et al., "Levels of Autonomy in AI Agent Systems," arXiv preprint arXiv:2502.15212 (2025).

4 Patel, Dwarkesh. "Why I Don't Think AGI Is Right Around the Corner." Dwarkesh Blog, June 2025.

5 Cihon, Peter, Chilling autonomy: Policy enforcement for human oversight of AI agents, 2024.

6 Kasirzadeh, Atoosa and Gabriel, Iason, Characterizing AI Agents for Alignment and Governance, April 2024; Soder, Lisa et al., An Autonomy-Based Classification: AI Agents, Liability and Lessons from the Automated Vehicles Act, Interface, April 2025; IBM, "What Are AI Agents?," accessed September 24, 2025; Hugging Face, "What is an Agent?", accessed September 24, 2025; Kapoor, Sayash et al., AI Agents That Matter, July 2024; Farmer, Harry and Smakman, Julia, Delegation Nation: Advanced AI Assistants and why they matter, Ada Lovelace Institute, February 2025; OWASP Agentic Security Initiative, Agentic AI – Threats and Mitigations, February 2025; Willison, Simon, "Using tweets gathered from https://til.simonwillison.net/twitter/collecting-replies," accessed September 24, 2025.

7 Kraprayoon, Jam et al., Agent Governance: A Field Guide, IAPS, April 2025; Kolt, Noam, Governing AI Agents, February 2025; McKenzie, Colleen et al., Research Agenda for Sociotechnical Approaches to AI Safety, March 2025; Toner, Helen et al., Through the Chat Window and Into the Real World: Preparing for AI Agents, Center for Security and Emerging Technology, October 2024; Cihon, Peter, Chilling autonomy: Policy enforcement for human oversight of AI agents, 2024. Oueslati, Amin and Staes-Polet, Robin, Ahead of the Curve: Governing AI Agents under the EU AI Act, The Future Society, June 2025; O'Keefe, Cullen et al., Law-Following AI: Designing AI Agents to Obey Human Laws, September 2025.

8 Howell, John and Ifayemi, Stephanie, Policy Alignment on AI Transparency, October 2024; Ifayemi, Stephanie et al., Decoding AI Governance: A Toolkit for Navigating Evolving Norms, Standards, and Rules, July 2025.

9 Mossberger, Karen et al., Virtual Inequality: Beyond the Digital Divide, 2003; Maslej, Nestor et al., The AI Index 2025 Annual Report, Institute for Human-Centered AI, Stanford University, April 2025; Ofcom, "Online Nation 2023 Report," November 2023; Störmer, Eckhard et al., Foresight – Using Science and Evidence to Anticipate and Shape the Future, Science for Policy Handbook, Ch. 12, 2020; Anthropic, "Economic Index"; Kasirzadeh, Atoosa and Gabriel, Iason, Characterizing AI Agents.

10 World Economic Forum, Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents, December 2024; Kraprayoon, Jam et al., Agent Governance; Kasirzadeh, Atoosa and Gabriel, Iason, Characterizing AI Agents.

11 Hammond, Lewis et al., Multi-Agent Risks from Advanced AI, Cooperative AI Foundation, February 2025; Bank of England Financial Policy Committee, Financial Stability in Focus: Artificial Intelligence in the Financial System, April 2025; Anthropic, Operating Multi-Client Influence Networks Across Platforms, April 2025; Bengio, Yoshua et al., International AI Safety Report 2025, February 2025; Slattery, Peter et al., The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence, August 2024; Kasirzadeh, Atoosa and Gabriel, Iason, Characterizing AI Agents; OWASP, Agentic AI – Threats and Mitigations.

12 Baldwin, Robert et al., Understanding Regulation: Theory, Strategy, and Practice, 2nd ed. 2012.

13 UK AI Security Institute, "Research Agenda", accessed September 24, 2025; Infocomm Media Development Authority, The Singapore Consensus on Global AI Safety Research Priorities: Building a Trustworthy, Reliable and Secure AI Ecosystem, May 2025.

14 Ranchordás, Sofia and Vinci, Valeria, Regulatory Sandboxes and Innovation-friendly Regulation: Between Collaboration and Capture, January 2024; Jeník, Ivo and Duff, Schan, How to Build a Regulatory Sandbox: A Practical Guide for Policy Makers, September 2020; Ministry of Science, Technology and Innovation, "National Technology and Innovation Sandbox (NTIS)" accessed September 24, 2025; Information Commissioner's Office, "The Guide to the Sandbox," accessed September 24, 2025; Financial Conduct Authority, Engagement Paper: Proposal for AI Live Testing, April 2025; EU AI Act, arts. 57, 60.

15 Chan, Alan et al., Infrastructure for AI Agents, June 2025; Dafoe, Allan, AI Governance: A Research Agenda, Centre for the Governance of AI, Future of Humanity Institute, August 2018; Pratt, Jacob and Tanjaya, Albert, Documenting the Impact of Foundation Models, Partnership on AI, February 2025.

16 Chan, Alan et al., Visibility into AI Agents, May 2024.

17 Chan, Alan et al., Infrastructure for AI Agents.

18 Chan, Alan et al., Infrastructure for AI Agents; Reuel, Anka et al., BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices, November 2024.

19 Ouimette, Marc E. et al., AI, Everywhere, All At Once: A New Policy Agenda for AI Success Through Faster Adoption, October 2024; Systma, Tobias, Managing AI's Economic Future: Strategic Automation Policy in an Era of Global Competition, RAND, May 2025.