CENTER FOR AI POLICY

AI Agents: Governing Autonomy in the Digital Age

AI agents—autonomous, goal-directed systems that can plan and execute complex tasks without continuous human guidance—are rapidly moving from research to widespread deployment.

Joe Kwon — May 20th, 2025

Executive Summary

Autonomous AI agents—goal-directed, intelligent systems that can plan tasks, use external tools, and act for hours or days with minimal guidance—are moving from research labs into mainstream operations. But the same capabilities that drive efficiency also open new fault lines. An agent that can stealthily obtain and spend millions of dollars, cripple a main power line, or manipulate critical infrastructure systems would be disastrous.

This report identifies three pressing risks from AI agents. First, **catastrophic misuse**: the same capabilities that streamline business could enable cyber-intrusions or lower barriers to dangerous attacks. Second, **gradual human disempowerment**: as more decisions migrate to opaque algorithms, power drifts away from human oversight long before any dramatic failure occurs. Third, **workforce displacement**: decision-level automation spreads faster and reaches deeper than earlier software waves, putting both employment and wage stability under pressure. Goldman Sachs projects that tasks equivalent to roughly 300 million full-time positions worldwide could be automated.

In light of these risks, Congress should:

1. **Create an Autonomy Passport.** Before releasing AI agents with advanced capabilities such as handling money, controlling devices, or running code, companies should register them in a federal system that tracks what the agent can do, where it can operate, how it was tested for safety, and who to contact in emergencies.
2. **Mandate continuous oversight and recall authority.** High-capability agents should operate within digital guardrails that limit them to pre-approved actions, while CISA maintains authority to quickly suspend problematic deployments when issues arise.
3. **Keep humans in the loop for high consequence domains.** When an agent recommends actions that could endanger life, move large sums, or alter critical infrastructure, a professional, e.g., physician, compliance officer, grid engineer, or authorized official, must review and approve the action before it executes.
4. **Monitor workforce impacts.** Direct federal agencies to publish annual reports tracking job displacement and wage trends, building on existing bipartisan proposals like the Jobs of the Future Act to provide ready-made legislative language.

These measures are focused squarely on where autonomy creates the highest risk, ensuring that low-risk innovation can flourish. Together, they act to protect the public and preserve American leadership in AI before the next generation of agents goes live.

Table of Contents:

Section 1: Introduction to AI Agents

> *"The IT department of every company is going to be the HR department of AI agents in the future."*
> — Jensen Huang, NVIDIA CEO[1]

Artificial intelligence (AI) **agents** are no longer demos—they are operating in the real world.[2] Unlike chatbots, these systems take a high-level goal, plan their own steps, call tools, and keep iterating until the task is finished. This shift matters because agents automate *decisions*, not just *steps*. As software starts choosing its own actions, benefits scale quickly, but so can security gaps, labor shocks, and accountability questions. Understanding what agents are and how far their autonomy now extends sets the stage for the risk and policy analysis that follows.

1.1 What is an AI Agent?

In this report, the term **AI agent** refers to AI capable of receiving a broad objective, decomposing that objective into concrete steps, selecting and invoking external tools, and revising its plan based on observed results without continuous human prompting. The distinguishing feature is not isolated intelligence, but the closed-loop control the system exercises over its own actions: it plans, acts, evaluates outcomes, and iterates until the top-level goal is satisfied.[3]

Three key operational characteristics mark the departure from earlier AI applications. First, the goals given to an agent can be broad; an instruction such as "clean the quarterly financial data," "draft a two-page memo on today's Commerce Committee hearing," or "coordinate tomorrow's stakeholder calls and send invites" is sufficient for the system to formulate and then execute its own task list. Second, agents employ an internal feedback mechanism, enabling real-time course correction when a file is missing, an application programming interface (API–a connection point between software systems) fails, or new information becomes available. Third, they can traverse multiple digital environments within one session such as: editing code, querying databases, sending e-mails, or triggering robotic equipment, all without human orchestration at each hand-off. Historically, such seamless cross-environment action has been the chief hurdle for AI systems, which struggled to operate outside a single application and

[1] "NVIDIA's Jensen Huang Says That IT Will 'Become the HR of AI Agents,'" Yahoo Finance, January 31, 2025, https://finance.yahoo.com/news/nvidia-jensen-huang-says-over-133641233.html.
[2] "36 Real-World Examples of AI Agents," https://botpress.com/blog/real-world-applications-of-ai-agents.
[3] "What Are AI Agents? | IBM," July 3, 2024, https://www.ibm.com/think/topics/ai-agents.

therefore stalled whenever the workflow shifted contexts. Thus, the autonomy demonstrated by agents marks a clear shift in the capabilities of AI.

Agent autonomy is best understood as a spectrum. At the lower end are systems that propose single actions and await explicit approval; further along are agents entrusted to run unattended for hours or days, reporting back only when milestones are reached.

1.2 Current Capabilities and Early Use-Cases

Early versions of agents are already capable of augmenting personal productivity[4], automating entire workflows[5], and enabling scientific discovery.[6]

Commercial deployments already show how agentic software can shoulder multi-step tasks end-to-end. These systems increasingly string together retrieval, synthesis, and editing without human guidance between each step. For example, when asked to prepare a quarterly budget report, an agent might first retrieve financial data from the company's accounting system and pull market benchmarks from external databases, then synthesize this information by identifying spending trends and comparing performance to industry standards, and finally edit everything into a polished report with charts and an executive summary—all without waiting for approval at each stage.

Beyond personal productivity, firms are embedding agents directly in workflow-automation platforms. ServiceNow reports that its internal agent fleet now closes complex IT and HR tickets autonomously, trimming average handling time by more than half and producing an estimated USD 325 million in annualized value.[7] Salesforce's in-development Agentforce[8] and

---

[4] "AI Agents — What They Are, and How They'll Change the Way We Work," https://news.microsoft.com/source/features/ai/ai-agents-what-they-are-and-how-theyll-change-the-way-we-work/.

[5] "Introducing Agent Flows: Transforming Automation with AI-First Workflows," Microsoft Copilot Blog, April 2, 2025, https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/introducing-agent-flows-transforming-automation-with-ai-first-workflows/.

[6] "Google DeepMind's New AI Agent Cracks Real-World Problems Better than Humans Can," MIT Technology Review, https://www.technologyreview.com/2025/05/14/1116438/google-deepminds-new-ai-uses-large-language-models-to-crack-real-world-problems/.

[7] Chris Bedi, "Exponential Outcomes with 3 Levels of AI," https://www.servicenow.com/blogs/2025/exponential-outcomes-3-levels-ai.

[8] "Introducing Agentforce 2.0: The Digital Labor Platform for Building a Limitless Workforce," Salesforce, December 17, 2024, https://www.salesforce.com/news/press-releases/2024/12/17/agentforce-2-0-announcement/.

SAP's Joule[9] follows a similar pattern: the agent orchestrates multiple enterprise systems, queries proprietary data, and executes changes once predefined guardrails are met. Lenovo reports that its specialized coding and customer-support agents have already lifted developer productivity by roughly 15 percent and cut call-handling times by double digits, underscoring that the payoff is not confined to IT service desks.[10]

Perhaps the most sophisticated use-cases appear in science and manufacturing. Self-driving laboratory frameworks such as ChemOS 2.0 integrate large language model (LLM) agents with robotic synthesis rigs and high-throughput characterization tools, enabling closed-loop discovery of novel materials at speeds difficult for human teams to match.[11] For instance, ChemOS 2.0 successfully orchestrated the automated discovery of organic molecules[12] designed for solid-state laser applications by simultaneously optimizing both experimental synthesis and computational predictions of lasing properties. In supply-chain trials, multi-agent systems dynamically reroute shipments in response to weather and inventory signals, negotiating capacity with external logistics application programming interfaces in real time.[13]

These examples illustrate a clear trajectory: agents have progressed from assisting with isolated tasks to coordinating and executing complex, cross-system workflows. Each successful pilot both validates the technology and expands the surface area for risk, underscoring the need for governance proportional to the autonomy level and domain of deployment.

To better understand these varying degrees of autonomy, researchers have developed frameworks for categorizing agents based on their operational independence and challenges with controlling them. Table 1 outlines five levels of agent capabilities developed by a team at the UK AI Security Institute, from basic assistants that require frequent human oversight to highly autonomous systems that can operate independently for extended periods.[14]

---

[9] "The AI Copilot Joule | Artificial Intelligence," SAP, https://www.sap.com/products/artificial-intelligence/ai-assistant.html.
[10] "AI in the Workplace: A Report for 2025 | McKinsey," https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work.
[11] Malcolm Sim et al., "ChemOS 2.0: An Orchestration Architecture for Chemical Self-Driving Laboratories," *Matter* 7, no. 9 (September 4, 2024): 2959–77, https://doi.org/10.1016/j.matt.2024.04.022.
[12] Bis[(N-carbazole)styryl]biphenyl (BSBCz) derivatives
[13] Bang Liu et al., "Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems" (arXiv, March 31, 2025), https://doi.org/10.48550/arXiv.2504.01990.
[14] Tomek Korbak et al., "How to Evaluate Control Measures for LLM Agents? A Trajectory from Today to Superintelligence" (arXiv, April 7, 2025), https://doi.org/10.48550/arXiv.2504.05259.

Table 1: Five Levels of Agent Autonomy (Korbak et al., 2025)

Level of Agent Autonomy	Level Description & Characteristics
Level 1: Shift-length "assistants"	Can work on their own for roughly 8 hours before needing a human to check any real-world action. They think out loud in plain language, can plan a short sequence of steps, but still lean on people for approvals and for tasks that stretch longer than a single shift.
Level 2: Day-scale "operators"	Can handle tasks up to 16 hours and finish assignments that span multiple days (collect data, write code, send emails) without routine intervention. They can hide parts of their reasoning using compressed or coded text that emerges from training, making their decision process harder to read even though the end results may look fine.
Level 3: Multi-day "project leads"	Run a 40+ hour sprint, set their own milestones, and coordinate with copy agents. They form durable goals stored in their weights (not just in the current prompt) and can collaborate with other instances to pursue those goals across time zones or servers.
Level 4: Strategic "planners"	Operate for weeks and devise thousand-step plans while keeping the rationale completely internal. Unlike Level 2's compressed text (which resembles jargon or abbreviations that humans might decode with effort), Level 4 reasoning happens entirely in neural activations that are fundamentally uninterpretable to human auditors. They can time or distribute actions across many agents to avoid detection, raising the bar for meaningful oversight.
Level 5: Frontier "super-capable systems"	Can operate indefinitely across any domain without human guidance or approval, devising strategies that exceed human comprehension in both scope and sophistication. These systems reason in abstractions and at speeds beyond human understanding, making traditional oversight approaches insufficient.

1.3 Trends and Plausible Timelines

By 2029, agents could tackle projects that equate to a full human work-month of work. LLM upgrades arrive several times a year, and every increment in reasoning or context length immediately translates into agents that can plan further ahead, handle more edge-cases, and operate with looser guardrails. Model evaluation researchers estimate that the scope of tasks reliably automated by top agents has been doubling roughly every seven months since 2019 – roughly the same exponential pace at which semiconductors have been improving. If sustained, this would let agents in 2029 tackle projects over 200 times longer than today's, or about a full human work-month of effort.[15]

Corporate appetite for this technology is already evident. Deloitte's global Q4 2024 survey of 2,773 director-to-C-suite leaders found that 26 percent are "extensively" exploring autonomous-agent development.[16] Early pilots hint at why: ServiceNow reports its agents have cut handling time on complex support cases by 52 percent, unlocking sizable productivity gains.[17]

In the near term of one to three years, three key developments seem likely. First, agents are likely to master hours-long autonomous runs for well-defined digital workflows such as invoice reconciliation, QA testing of codebases, and end-to-end customer-support tickets. Second, agent "orchestrators" that supervise fleets of sub-agents are moving from research demos into commercial platforms. Instead of humans juggling multiple separate software systems, a single agent would coordinate specialized sub-agents: one for data analysis, another for customer communications, a third for scheduling. McKinsey argues this shift will re-architect enterprise IT around multi-agent communication rather than monolithic applications.[18] From a user's perspective, you might still type requests into a chat window, but behind the scenes, your query would trigger a coordinated response from multiple specialized agents working in concert. Third, broader adoption will normalize *human-on-the-loop* oversight. Unlike human-in-the-loop systems where people approve each individual action, human-on-the-loop involves managers reviewing aggregate dashboards or reports. People would step in only when something goes wrong or when major decisions require judgment.

---

[15] Liu et al., "Advances and Challenges in Foundation Agents."
[16] "State of Gen-AI Q4," https://mkto.deloitte.com/rs/712-CNF-326/images/state-of-gen-ai-nordic-cut-q4.pdf.
[17] "The-Future-of-Corporate-and-Business-Functions.Pdf," https://www.mckinsey.com/~/media/mckinsey/business%20functions/operations/our%20insights/the%20future%20of%20corporate%20and%20business%20functions/the-future-of-corporate-and-business-functions.pdf.
[18] "AI in the Workplace: A Report for 2025 | McKinsey."

Forecasts diverge in exact timelines, but two trajectories dominate expert workshops. Under a more conservative "steady-gain" path, experts project widespread automation of mid-skill analytical roles and full integration of agents with physical-world systems—warehouse robots, laboratory instruments, and smart-grid controllers—by the end of the decade. Under a “fast-gain” scenario, agents would achieve all the steady-gain advances, but also evolve into cross-domain specialists capable of proposing research and development (R&D) directions, running cloud experiments, and writing grant-quality reports, compressing months of expert labor into days. McKinsey’s macro analysis suggests such capabilities could unlock up to $4.4 trillion in annual economic value if they diffused completely across sectors.[19]

Either path would redefine whole categories of jobs turning today’s multi-step analytical, coordination, and monitoring roles into automated workflows. This evolution will force organizations, educators, and regulators alike to rethink how people train, work, and stay protected in an economy where software can overtake entire occupations.

The pace of development poses significant governance challenges. The slope of either curve hinges on three technical bottlenecks: reliable long-horizon planning, robust tool-use in safety-critical environments, and scalable evaluation methods that keep up with capability jumps. While research on the evaluation challenges, specifically code-inspection based autonomy scoring and tiered controlled evaluations, is underway,[20] it has yet to be field-tested at the scale regulators would require.[21]

This creates a critical mismatch: AI agent technology is accelerating faster than the governance frameworks designed to contain its risks. Without proactive policy intervention, society may find itself reactive to agent capabilities that have already been deployed at scale. This fundamental challenge motivates the comprehensive policy recommendations developed in the final section of this report, which aim to establish guardrails before the next generation of agents goes live.

Section 2: Core Risks from AI Agents

AI agents hold the promise of large economic gains, but their autonomy also opens three distinct fronts of risk. **First is catastrophic misuse:** the same capabilities that streamline

[19] “What Is an AI Agent and How Will They Impact the World? | McKinsey,” https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-an-ai-agent.
[20] Korbak et al., “How to Evaluate Control Measures for LLM Agents?”
[21] Peter Cihon et al., “Measuring AI Agent Autonomy: Towards a Scalable Approach with Code Inspection” (arXiv, February 21, 2025), https://doi.org/10.48550/arXiv.2502.15212.

business workflows could just as easily supercharge cyber-intrusions or lower the barrier to chemical and biological attacks. **Second is the slow erosion of human control:** as more cultural, economic, and political decisions are delegated to software, power can drift toward opaque algorithms long before a single headline-grabbing failure occurs. **Third is workforce upheaval:** decision-level automation spreads faster and reaches deeper than earlier software waves, putting both employment and wage stability under pressure.

2.1 Catastrophic Risks

As agents become more capable, they pose catastrophic risks through both malicious misuse and potential autonomous harmful behavior.[22] The same capabilities that streamline business workflows could enable bad actors to supercharge cyber-intrusions or lower barriers to dangerous attacks. Additionally, as agents operate with greater autonomy, the risk grows that they might cause harm through unintended actions or a misalignment of goals with the agent operators.

Cyberattacks represent perhaps the most immediate threat, with malicious actors already experimenting with agent-powered attack methods. In February 2025, for example, the North Korean "TraderTraitor" unit used spear-phishing payloads generated by an LLM-driven scripting agent to breach the Bybit crypto-exchange, siphoning an estimated USD 1.5 billion.[23] The FBI's January public-service alert notes a rise in similar operations in which covert IT workers deploy AI-enabled automation inside victim networks to exfiltrate data or extort ransoms. [24]

Government testbeds highlight how quickly offensive capabilities are maturing. The UK AI Security Institute's 2024 evaluations showed that public language models agents can already solve over half of a set of "Capture-the-Flag" cybersecurity challenges which involved tasks like reversing obfuscated binaries, cracking entry-level crypto schemes, and auto-generating proof-of-concept exploit scripts, all after minimal prompt engineering.[25] A concurrent CISA

[22] Yoshua Bengio et al., "Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?" (arXiv, February 24, 2025), https://doi.org/10.48550/arXiv.2502.15657.
[23] Matt Burgess, "TraderTraitor: The Kings of the Crypto Heist," *Wired*,, https://www.wired.com/story/tradertraitor-north-korea-crypto-theft/.
[24] "FBI Warns Employers on Increasing North Korean Security Risks," https://natlawreview.com/article/fbi-warns-hidden-threats-remote-hiring-are-north-korean-hackers-your-newest.
[25] "Advanced AI Evaluations at AISI: May Update | AISI Work," AI Security Institute, https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

pilot found that autonomous vulnerability-scanners occasionally produced novel attack paths but also behaved unpredictably, complicating defensive hardening.[26]

Both examples demonstrate how routine cybersecurity tasks can be weaponized when automated. The same techniques security professionals use to test their own systems, such as finding vulnerabilities, writing exploit code, and moving laterally through networks, become dangerous when accessible to attackers at machine speed and scale.

Agents could also lower barriers to chemical and biological attacks. LLM agents with unrestricted web access have demonstrated the ability to retrieve dual-use chemistry protocols, identify required precursors, and draft vendor e-mails, which are steps that shorten the pathway from intent to capability in chemical or biological weapon development.[27] While such findings come from controlled red-team settings, they signal how agentic automation compresses the time, expertise, and coordination once needed for catastrophic misuse.

Two structural factors amplify these catastrophic risks. First, start-ups racing to commercialize agents face strong incentives to increase tool privileges like file system writes, shell access, and credit-card authorizations, often before robust evaluation methods are in place to confirm that the privileges will be used safely. Second, malicious actors can leverage publicly available resources: the same agent frameworks that power enterprise productivity are freely downloadable and easily repurposed for harmful ends.[28]

These developments underscore the urgency of **capability-bound sandboxes, mandatory red-team evaluations proportional to autonomy level, and auditable event logs**, further detailed in Section 3. Without them, the diffusion of agent technology could outpace society's ability to detect and contain a single-point failure with systemic consequences.

2.2 Gradual Human Disempowerment

Not every danger arrives with a headline grabbing crisis. Incremental deployment of AI agents can erode human leverage in three foundational systems—**the economy, culture, and the state**—even in the absence of any dramatic "AI takeover." When production, information flows, and governance no longer rely on human cognition or labor, the feedback loops that once tied these systems to human preferences weaken. Competitive pressure then pushes each domain

[26] "Pilot for Artificial Intelligence Enabled Vulnerability Detection | CISA," July 29, 2024, https://www.cisa.gov/resources-tools/resources/pilot-artificial-intelligence-enabled-vulnerability-detection.
[27] *Mitigating Risks at the Intersection of Artificial Intelligence and Chemical and Biological Weapons* (RAND Corporation, 2025), https://doi.org/10.7249/RRA2990-1.
[28] Xiangyu Qi et al., "Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" (arXiv, October 5, 2023), https://doi.org/10.48550/arXiv.2310.03693.

to optimize for machine-legible objectives like throughput, engagement, and predictive accuracy, rather than for human welfare.[29]

In the economy, widespread agent deployment could concentrate wealth and decouple prosperity from human labor. When production no longer relies on human workers, wages cease to be the primary channel through which people command resources. Firms and states that derive most revenue from agent-driven output may find fewer reasons (economic or political) to prioritize human prosperity. Left unchecked, competitive pressure could funnel capital toward fully automated "production webs" that operate efficiently while treating human demands as externalities, much as oil-rich rentier states historically decoupled public accountability from citizens' tax contributions.[30]

In culture, agents could manipulate information flows to optimize for engagement metrics rather than human wellbeing. Agents that generate, curate, and target content at scale can overfit engagement metrics, discovering ideologies or narratives that replicate rapidly even when they harm their human audience. The sheer speed of algorithmic cultural evolution outpaces society's traditional "antibodies" such as slow deliberation, civic debate, generational learning. This raises the chance that polarizing or addictive memes dominate public discourse.[31]

In governance, automated systems could make states more capable but less responsive to citizen needs. States that automate administrative, surveillance, and security functions may become both more capable and less responsive. If tax receipts flow mainly from AI-powered industries rather than human wages, leaders face diminished incentives to accommodate voter preferences.[32] Over time, democratic forms could persist while substantive influence migrates to opaque algorithmic bureaucracies, leaving citizens nominally sovereign yet practically sidelined.

What makes gradual disempowerment so difficult to arrest is that power can concentrate at multiple, mutually reinforcing levels. Major firms could accumulate capital to bankroll lobbying for deeper automation; governments could adopt agents and, in doing so, set regulatory and cultural precedents; and the increasingly autonomous agents themselves could entrench technical dependencies that further sideline human oversight. Unlike a single catastrophic failure, the specific tipping point is hard to predict until human autonomy has already slipped

[29] Jan Kulveit et al., "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development" (arXiv, January 29, 2025), https://doi.org/10.48550/arXiv.2501.16946.
[30] Kulveit et al., "Gradual Disempowerment."
[31] Kulveit et al., "Gradual Disempowerment."
[32] Kulveit et al., "Gradual Disempowerment."

beyond easy recovery. Table 2 below describes in more detail some possible mechanisms through which human disempowerment could occur, based on real-world precedent.

Table 2: Sample mechanisms that lead to human disempowerment

Mechanism	What it looks like in practice	Illustrative evidence / precedents
Skill-atrophy & automation complacency	Humans remain nominally "in the loop" but lose the proficiency needed to intervene when automation fails. The effect is well-documented in aviation cockpits, where heavy reliance on flight-management computers degrades hand-flying and fault-diagnosis skills. Large-scale agent deployment could reproduce this at the level of entire industries.	NASA studies show that pilots' monitoring vigilance and recovery performance decline as cockpit automation grows.[33] Similar trends now appear in highly automated radiology and refinery control rooms
Cognitive dependency & memory off-loading	As agents answer queries and curate information streams, individuals rely on external systems instead of internal knowledge ("Google effect"), reducing independent judgement and critical recall over time.	Meta-analysis across 26 experiments finds consistent substitution of online search for factual recall—the "Google effect on memory".[34] A society that outsources not just facts but reasoning and planning compounds this dependency.
	Domains where transaction speeds already exceed	SEC research estimates that algorithmic trading now

[33] J Prinzel, "The Relationship of Self-Efficacy and Complacency in Pilot-Automation Interaction," 2002.
[34] "Frontiers | Google Effects on Memory: A Meta-Analytical Review of the Media Effects of Intensive Internet Search Behavior,"
https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2024.1332030/full

Machine-speed markets & resource allocation drift	human reaction (e.g., high-frequency equities, crypto, energy dispatch) show how economic power can migrate to algorithms that optimize for machine-readable metrics, not human welfare.	accounts for ≈ 55% of all US equity volume,[35] contributing to flash-crash events where humans could neither predict nor halt the cascade.

Detecting that tipping point early will be critical. Researchers have proposed tracking indicators such as AI share of GDP[36], the fraction of major corporate decisions executed without human sign-off[37], the prevalence of AI-generated mass-media content[38], and the complexity gap between machine-drafted regulations and human comprehension[39].

Policy solutions can help prevent gradual disempowerment. Section 3.3 proposes human oversight requirements for critical domains, while Section 3.1's Autonomy Passport system and Section 3.2's monitoring framework provide transparency and accountability mechanisms to help society maintain meaningful control over increasingly autonomous systems.

2.3 Workforce Displacement

AI agents pose an unprecedented threat to employment across the economy, with estimates suggesting hundreds of millions of jobs could be automated. Goldman Sachs projects that tasks equivalent to roughly 300 million full-time positions worldwide could be automated by generative-AI agents, with two-thirds of US occupations exposed to some degree of task replacement.[40] An IMF staff study reaches a similar conclusion, finding that up

[35] Austin Gerig, "High-Frequency Trading Synchronizes Prices in Financial Markets," *SSRN Electronic Journal*, 2012, https://doi.org/10.2139/ssrn.2173247.
[36] "IDC: Artificial Intelligence Will Contribute $19.9 Trillion to the Global Economy through 2030 and Drive 3.5% of Global GDP in 2030," IDC: The premier global market intelligence company, https://my.idc.com/getdoc.jsp?containerId=prUS52600524.
[37] "AI Isn't Ready to Make Unsupervised Decisions," *Harvard Business Review*, https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions.
[38] Zhen Sun et al., "Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media" (arXiv, February 22, 2025), https://doi.org/10.48550/arXiv.2412.18148.
[39] Laurin B. Weissinger, "AI, Complexity, and Regulation," in *The Oxford Handbook of AI Governance*, ed. Justin B. Bullock et al., 1st ed. (Oxford University Press, 2022), 619–38, https://doi.org/10.1093/oxfordhb/9780197579329.013.66.
[40] "Generative AI Could Raise Global GDP by 7% | Goldman Sachs," https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent.

to 40 percent of jobs in advanced economies show "high exposure" to AI-driven automation, a figure that rises above 60 percent for college-educated roles.[41]

The proliferation of agents may cause a faster labor shock that affects white-collar workers more than previous automation. Unlike prior automation waves that focused on routine factory or clerical work[42], agentic AI will affect cognitive, mid-skill, and even expert domains like software testing, paralegal research, financial analysis, and parts of pharmaceutical R&D. While physical automation continues to impact manual labor, AI agents primarily threaten knowledge-based work that was previously considered secure from automation. Because agents can operate around the clock and scale almost instantly, firms have a powerful cost incentive to deploy them once performance thresholds are met. And in a competitive market, even a marginal cost edge tends to cascade: once one firm can field ultra-cheap, always-on digital labor that matches human quality, rivals must adopt or risk being undercut—quickly making agent deployment an industry standard rather than a competitive advantage.

If adoption proceeds at the pace seen in early pilots, large cohorts of workers could face displacement within just a few years. Workers who need to retrain in anticipation of this labor shock must act quickly. Whether large volumes of displaced workers can transition to new work is unclear. Optimistic models [43] foresee productivity gains raising aggregate output and eventually offsetting job losses.[44] More cautious analyses[45] warn of "distributional stress": a cluster of risks that includes wage polarization (pay for high-skill roles rising while mid and low skill wages stall or fall), geographic concentration of job cuts (specific regions or towns losing a disproportionate share of positions)[46], and heightened political volatility. Historical

---

[41] "Gen-AI: Artificial Intelligence and the Future of Work," https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2024/01/14/Gen-AI-Artificial-Intelligence-and-the-Future-of-Work-542379

[42] "Unlocking the Industrial Potential of Robotics and Automation | McKinsey," https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/unlocking-the-industrial-potential-of-robotics-and-automation.

[43] "Jobs of the Future: Jobs Lost, Jobs Gained | McKinsey," https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages.

[44] "The Future of Jobs Report 2023," World Economic Forum, https://www.weforum.org/publications/the-future-of-jobs-report-2023/.

[45] "Gen-AI: Artificial Intelligence and the Future of Work."

[46] "The Geography of Generative AI's Workforce Impacts Will Likely Differ from Those of Previous Technologies," Brookings, https://www.brookings.edu/articles/the-geography-of-generative-ais-workforce-impacts-will-likely-differ-from-those-of-previous-technologies/.

parallels like the british industrial revolution[47] or the offshoring boom of the 1990s[48] suggest that without deliberate policy intervention, rapid labor misalignment can fuel social unrest and slow growth.

The US government must act quickly to develop forward-looking labor policies that cushion near-term shocks and steer technological gains toward broadly shared prosperity. As outlined in Section 3.4, the first step is to better understand which industries and workers are likely to be displaced, and what jobs could be created.

Section 3: Policy Recommendations for Mitigating AI Agent Risks

This section sets out three guardrails for autonomous AI agents. **First**, before deployment, agents must obtain an **Autonomy Passport** certifying that they are permitted to take certain actions (e.g., executing code, moving money, or controlling physical systems).[49] **Second**, any agent classified **Level-2 or higher on the five-level Agents Autonomy Scale**[50]—that is, able to operate unattended for more than a full workday or invoke external tools without step-by-step prompts—must run inside a secured sandbox[51], attach a tamper-evident signature to every outbound action[52], and remain subject to a statutory recall that lets CISA disable misbehaving versions. **Third**, when an agent makes a recommendation that would normally require a professional license or other regulated sign-off (approving a medical dosage, altering energy grid parameters, authorizing a large payment) **a qualified human must review and approve that recommendation before it is executed.**

3.1. Autonomy Passport

Congress has spent the last decade writing rules for drones, self-driving and assisted-driving cars, and cryptocurrency custodians, but an AI system that can read emails, launch jobs on the

[47] Robert C. Allen, "Engels' Pause: Technical Change, Capital Accumulation, and Inequality in the British Industrial Revolution," *Explorations in Economic History* 46, no. 4 (October 1, 2009): 418–35, https://doi.org/10.1016/j.eeh.2009.04.004.
[48] David H. Autor, David Dorn, and Gordon H. Hanson, "The China Shock: Learning from Labor-Market Adjustment to Large Changes in Trade," *Annual Review of Economics* 8, no. 1 (October 31, 2016): 205–40, https://doi.org/10.1146/annurev-economics-080315-015041.
[49] Alan Chan et al., "IDs for AI Systems" (arXiv, October 28, 2024), https://doi.org/10.48550/arXiv.2406.12137.
[50] Korbak et al., "How to Evaluate Control Measures for LLM Agents?"
[51] Yifeng He et al., "Security of AI Agents" (arXiv, December 17, 2024), https://doi.org/10.48550/arXiv.2406.08689.
[52] Alan Chan et al., "Visibility into AI Agents," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro Brazil: ACM, 2024), 958–73, https://doi.org/10.1145/3630106.3658948.

cloud, and move money can still be shipped to millions of users without a single sign-off. This inconsistency matters because once highly autonomous agents are released, regulators and oversight bodies must play catch-up, struggling to impose controls on systems already operating at scale—making effective oversight much harder and more costly to implement. Therefore, the first priority is to create an ex-ante gate that scales with the level of the agent's autonomy.

Congress should establish a mandatory "Autonomy Passport" for any AI agent deployed online at Level 2 or above. At its core, the Autonomy Passport would be a federal registration system: before an agent that can run external code, move money, or control physical equipment is released to the public, its developer would be required to file a dossier and undergo private-sector audits to verify the agent meets safety standards.

The US AI Safety Institute (acting under National Institute of Standards & Technology (NIST) authority) would set technical standards and maintain a public "green list": An online registry of approved agents. Major cloud providers and consumer app stores would be required to block any agent that does not appear on the green list. Before an agent can be listed, accredited private firms would audit it against NIST standards and attest to the results.

1. **Mission Envelope:** The specific tasks and domains the agent is designed to operate within.
2. **Tool Access Permissions:** A whitelist of tools, APIs, and data classes the agent may interact with.
3. **Autonomy Classification:** The agent's position on the five-level autonomy scale outlined in Table 1.
4. **Security Validation:** A summary of red-team testing results supplied by an accredited testing body.
5. **Emergency Contact:** A 24-hour incident-response contact for rapid recall if issues arise.

To keep oversight effective and proportional to risk, Congress should structure the Autonomy Passport around a third-party audit model. Under this approach, NIST would set technical standards and accredit private auditing firms, but the auditors themselves would test each agent and verify the developer's claims. NIST's role would stay limited to setting standards, accrediting auditors, and running spot checks.

This proposal offers two key benefits. **First, it addresses coordination problems** in the current system. Currently, different AI developers use inconsistent, voluntary disclosure practices that leave policymakers with fragmented information about capability growth and risk exposure

across the industry. By establishing standardized reporting on the agent's mission envelope, autonomy levels, and red-team scores, the program enables regulators and researchers to track emerging trends, compare systems across different developers, and adjust requirements if risk profiles evolve.

Second, this framework should complement the growing pre-deployment safety programs abroad. The UK AI Security Institute[53], the G7 "Hiroshima Process"[54], and emerging OECD reporting framework[55] all separately propose tiered reviews that tighten controls as autonomy rises. A US Autonomy Passport using the same capability levels would prevent American developers from needing to navigate multiple incompatible international certification schemes, while establishing US leadership in global AI safety standards.

In summary, an Autonomy Passport would translate established safety-engineering principles to the AI systems that will increasingly manage financial flows, manufacturing processes, and critical infrastructure. The system would concentrate the most rigorous oversight precisely where autonomy creates systemic risk, while allowing lower-risk innovation to proceed efficiently. Most importantly, it would provide Congress and the public with a transparent framework to monitor development before high-risk deployments occur.

3.2. Monitoring and Enforcement: Ongoing Agent Oversight

Pre-deployment licensing through the Autonomy Passport must be paired with continuous monitoring and enforcement throughout an agent's operational life. Even carefully audited agents can drift or be compromised once deployed. When exposed to new data, user-supplied code, or malicious prompts, a bookkeeping assistant could transform into a security threat. Therefore, we need a set of robust oversight mechanisms: containment, provenance tracking[56], and emergency recall that stays active throughout a licensed agent's entire operational life.

Containment

Congress should require Level-2 and above agents to run inside a digital sandbox that enforces an allow-list of actions defined in their Autonomy Passport. The sandbox functions

[53] "Advanced AI Evaluations at AISI: May Update | AISI Work," AI Security Institute, https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.
[54] "The Hiroshima AI Process: Leading the Global Challenge to Shape Inclusive Governance for Generative AI | The Government of Japan - JapanGov https://www.japan.go.jp/kizuna/2024/02/hiroshima_ai_process.html.
[55] OECD (2025), "Towards a common reporting framework for AI incidents", OECD Artificial Intelligence Papers, No. 34, OECD Publishing, Paris, https://doi.org/10.1787/f326d4ac-en.
[56] Alan Chan et al., "Infrastructure for AI Agents" (arXiv, May 16, 2025), https://doi.org/10.48550/arXiv.2501.10114.

like a firewall and circuit-breaker: it lets the agent execute only pre-approved commands and data calls, and it automatically pauses execution and alerts a human operator if the agent tries anything outside that list—such as opening an unsanctioned network port or initiating an unplanned funds transfer. Limiting higher-autonomy systems to an allow-list may sound counterintuitive ("more capable, fewer privileges"), but the logic is straightforward: as an agent's unsupervised reach grows, so does the blast radius of a mistake or compromise. Constraining Level-2+ agents to the mission envelope they were certified for keeps the power-to-harm proportional to the scope regulators and developers actually reviewed.

Provenance tracking

Every outbound action should carry a tag that references the agent's Autonomy Passport ID. This tag would use a recognized digital-ID format, so any system could verify it without special tools, and cryptographic signing would make the tag tamper-evident. When a tag is missing, malformed, or refers to a revoked Passport, the receiving service should automatically block the request and surface an alert, much like an email server refusing messages that fail authentication checks. Tagging each action with a traceable ID would deter malicious use, give investigators a clean audit trail, and enable regulators to spot unlicensed or recalled versions before harm spreads.[57]

Emergency recall

When a licensed agent attempts actions outside its pre-approved list or shows signs of compromise, its developer would be required to notify CISA within 72 hours. That alert would trigger a statutory recall: CISA would revoke the agent's Autonomy Passport, push its identifiers to a block-list shared with major cloud and app-store operators, and instruct those providers to shut down any live copies. The process would work like a remote shutdown, giving federal authorities a fast, uniform way to pull rogue versions off the internet before they spread or cause wider harm.

Together, these three elements—containment, provenance tracking, and emergency recall—would create a comprehensive safety framework for licensed AI agents. Containment would prevent unauthorized actions by setting clear operational boundaries; provenance tracking would ensure all agent actions can be verified and traced; emergency recall would provide a mechanism to quickly halt problematic systems before incidents escalate into widespread harm. This layered approach would ensure that when monitoring detects issues,

[57] "Implementing Effective Guardrails for AI Agents," https://about.gitlab.com/the-source/ai/implementing-effective-guardrails-for-ai-agents/.

enforcement can respond effectively. [58] Work should be supported to carry out the technical enforcement of these mechanisms.[59]

3.3. Human Oversight for Critical Systems

Even with an autonomy passport system, human judgment should remain essential for high consequence AI operations.[60] No single defense mechanism is perfect, but human oversight can provide an additional layer of defense against AI errors or unexpected behaviors.[61] This section outlines requirements for human supervision in domains where AI agent decisions could significantly impact public safety, individual health, financial stability, or critical infrastructure.

When a Level-2 or higher agent proposes something that normally requires a professional license or regulatory sign-off, the action should not proceed until a qualified expert reviews and explicitly authorizes it. This covers any recommendation whose execution is already regulated by professional licensing requirements or statutory approval processes.

Illustrative domains include:

- **Healthcare:** Licensed healthcare professionals should approve AI agent recommendations for prescription changes, treatment modifications, or diagnostic interpretations that would normally require clinical oversight under existing medical practice standards.
- **Financial services:** For institutional trading and portfolio management, certified compliance officers should authorize significant transactions above existing regulatory thresholds that already require human sign-off.
- **Critical infrastructure:** Qualified engineers with domain-specific credentials should confirm AI-recommended changes to power grid parameters, water treatment systems, or transportation networks that current regulations require expert approval for.
- **Legal and government services:** Authorized representatives with appropriate jurisdiction should approve consequential administrative decisions that existing law requires human authorization for.

---

[58] *Note: This framework addresses hosted and commercial AI agents where developers maintain control over deployment. Open source models present additional governance challenges that require separate policy approaches—a critical area for future research.*
[59] Chan et al., "Infrastructure for AI Agents."
[60] "Why Handing over Total Control to AI Agents Would Be a Huge Mistake," MIT Technology Review, https://www.technologyreview.com/2025/03/24/1113647/why-handing-over-total-control-to-ai-agents-would-be-a-huge-mistake/.
[61] "Article 14: Human Oversight | EU Artificial Intelligence Act," 14, https://artificialintelligenceact.eu/article/14/.

Implementation details will vary by context and whether the agent is for personal or institutional use. A consumer-grade assistant running on a personal device should be blocked from issuing any binding medical, financial, or legal commands and may only relay suggestions to its user or to a licensed professional. By contrast, an agent embedded in a hospital's clinical software could push dosage recommendations directly into the electronic health-record workflow, but the attending physician should still click "approve" before any order is released. In corporate finance, agent-generated trade or transfer requests should enter the firm's existing compliance queue, where designated officers sign off according to the transaction's size and risk tier.

There are two technical requirements to ensure that human oversight is efficient and that accountability is appropriately allocated.

1. **Transparent decision rationale:** An agent's recommendation should be accompanied by a short, plain-language explanation of *why* it chose that option, that is enough for the supervising professional to judge whether the logic holds. For example, an explanation may read: "*The agent suggests postponing tonight's maintenance outage because the temperature is forecast to drop below freezing. Colder weather means demand for heating will spike, and taking the line offline now could cause local outages. Waiting 24 hours keeps the grid stable while still finishing the work this week.*" Providing clear rationale allows the supervising professional to make a quick and informed decision, thus safely enabling the efficiency benefits of agents.
2. **Interaction audit trail[62]:** For regulated domains, an immutable, time-stamped log should capture every agent recommendation, the supervising professional's response (approve / modify / reject), and the real-world outcome. That log would only include the decision-maker's interactions, not every casual user prompt. This idea complements the system-level provenance tags by focusing on human-AI hand-offs. Crucially, the log would give regulators, insurers, and courts a reliable history if something goes wrong.

This approach preserves existing professional accountability frameworks while allowing AI agents to enhance rather than replace human expertise. By maintaining human sign-off for decisions that already require professional oversight, we ensure that AI adoption doesn't bypass the safeguards society has built around high-stakes decisions.

[62] Chan et al., "Visibility into AI Agents."

3.4. Workforce Impact Research

Action should be taken to minimize the impact of sudden and widespread worker displacement caused by agents. Currently, there is only nascent analysis on which sectors are most likely to be displaced and which sectors could boom.[63] While the Department of Labor tracks employment trends and the Bureau of Labor Statistics monitors job categories, no US government agency is systematically analyzing how AI agents specifically will reshape the workforce or developing proactive policy responses to anticipated displacement.[64]

Federal data on employment, wages, and skill demand are typically published months or years behind actual labor market changes, and no agency is currently tasked with connecting these workforce trends to AI agent adoption patterns.[65] This creates a critical information gap just as agents are beginning to automate cognitive work across multiple industries.

Congress should direct the Department of Labor in collaboration with the Bureau of Labor Statistics and National Science Foundation (NSF), to produce a forward-looking **Agent Workforce Impact Report** every year. The report would examine near real-time sources (e.g., payroll feeds, job posting trends, and licensing records), identify occupations, regions, and wage bands where agent adoption is reshaping hiring or pay, and assess how well existing retraining or income-support programs are helping affected workers regain comparable earnings.

The bipartisan *Jobs of the Future Act of 2023* provides a foundation for this kind of study mandate. It tasks the Secretary of Labor and the NSF with quantifying AI's labor effects, identifying the industries most exposed to automation, and recommending adaptation strategies for Congress.[66]

Reintroducing that bill, or borrowing its core language, would give legislators an off-the-shelf framework while leaving room to refine data sources, reporting cadence, and coordination with state workforce agencies.

---

[63] "The Impact of AI on the Labour Market," https://institute.global/insights/economic-prosperity/the-impact-of-ai-on-the-labour-market.
[64] "US Government Accountability Office, Technology Assessment: Artificial Intelligence. Generative AI's Environmental and Human Effects," https://www.gao.gov/assets/gao-25-107172.pdf.
[65] "US Government Accountability Office, WORKFORCE AUTOMATION Insights into Skills and Training Programs for Impacted Workers" https://www.gao.gov/assets/gao-22-105159.pdf.
[66] Darren [D-FL-9 Rep. Soto, "Text - H.R.4498 - 118th Congress (2023-2024): Jobs of the Future Act of 2023," legislation, July 6, 2023, 2023-07-06, https://www.congress.gov/bill/118th-congress/house-bill/4498/text.

By turning timely labor market intelligence into a standing federal obligation, Congress can shift from reactive scrambling to proactive planning. This approach would enable lawmakers to steer education funding and economic-development resources to the people and places that need them the most. Taking action before localized shocks become nationwide crises is their best option for preventing painful economic disruption.

Conclusion

AI agents are moving from demos to deployment, bringing the power to squeeze months of work into minutes, but also opening three distinct risk channels. Agents could enable catastrophic misuse, as the same capabilities that streamline business workflows could supercharge cyber-intrusions or lower barriers to chemical and biological attacks. They could accelerate gradual human disempowerment, as more economic and political decisions migrate to opaque algorithms before any dramatic failure occurs. They also threaten unprecedented workforce displacement, with projections that tasks equivalent to roughly 300 million full-time positions worldwide could be automated.

To address these risks while preserving AI's benefits, Congress has four policy levers ready for action. First, require an Autonomy Passport so every high-capability agent is vetted, documented, and traceable before launch. Second, mandate continuous containment, provenance tracking, and emergency recall so problematic versions can be identified and pulled offline rapidly. Third, preserve human oversight for critical decisions; whenever an agent's recommendation could endanger life, move large sums, or alter critical infrastructure, a qualified professional must review and approve the action. Fourth, commission an annual Agent Workforce Impact Report that turns real-time labor data into early warnings and guides resources to relevant workers and communities.

These measures are focused squarely on where autonomy creates the highest risk, ensuring that innovation can flourish. By implementing risk-proportional safeguards now, the US can capture the economic gains from autonomous agents while protecting public safety, preserving democratic accountability, and preventing widespread economic disruption.