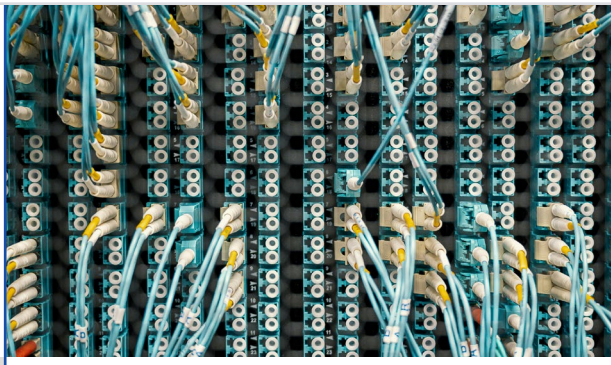
Google

May 2026

The TRUST Framework

A Policy Approach to Agentic AI



01 Executive Summary

Executive Summary

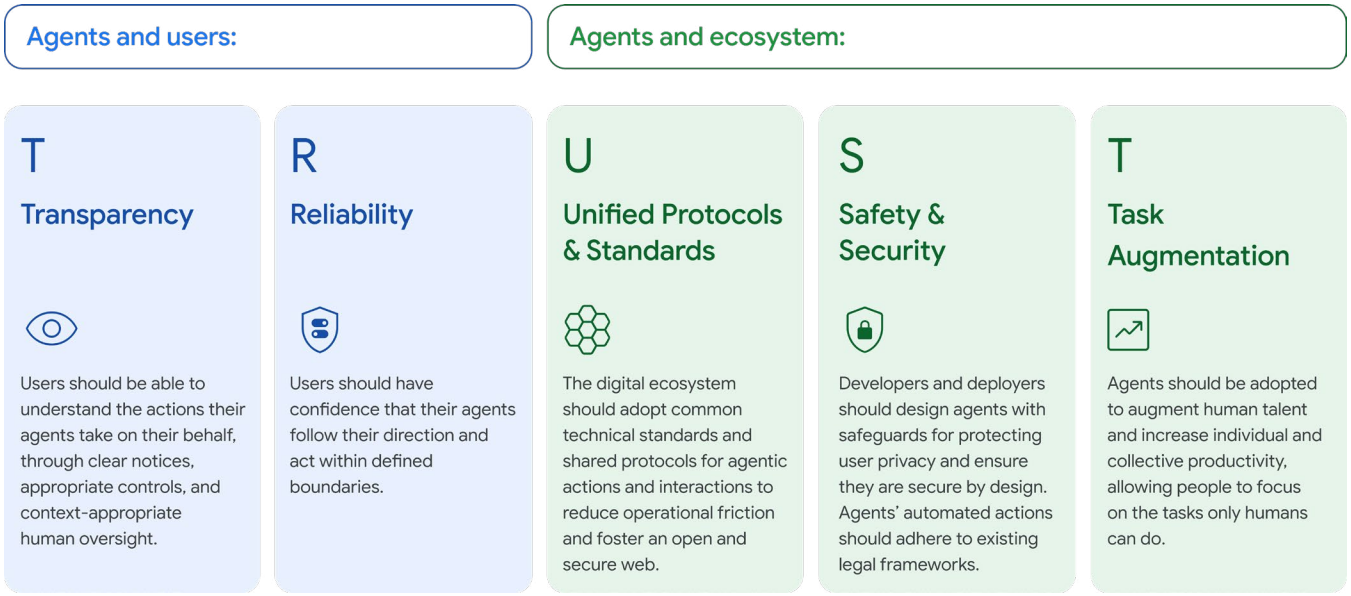
AI agents are the next step in the evolution of AI technology. Building on prior AI advances in reasoning and remembering context, and adding the ability to use digital tools, agents move beyond chatbots to plan and take action on a user's behalf and at their direction. These new capabilities are poised to change the ways people use computers and engage online, creating new connections among users, businesses, and digital infrastructure. Agents promise to not only help relieve people from tedious and time-consuming tasks but also help advance how we solve complex problems in important fields ranging from healthcare to energy and disaster response.

As AI moves from "answering" to "acting," it raises new questions—and renews some existing questions—about privacy, safety, and security, and AI's impacts on broader systems like digital commerce and the labor market. On the one hand, agents process tasks at a speed and scale that could raise issues for website traffic, technical permissions, and digital protocols, and can influence what information and services users see. On the other hand, agents can accelerate entrepreneurship, increase commercial transactions, and produce unprecedented gains in productivity, potentially unlocking trillions of dollars in economic value. This new economic activity may create entirely new categories of jobs, and spur changes in how current jobs get done, and perhaps catalyze larger shifts in economic and digital ecosystems.

Because we are only at the early stages of this technology, we don't know exactly what will change and what will stay the same. Given this uncertainty, regulatory approaches should prioritize continuity over added complexity. Proven legal frameworks provide continuity and security and avoid regulatory fragmentation. But we can't stop there. We need quick and nimble industry standards around governance, oversight, safety, and security best practices to manage agentic capabilities and potential risks. The rise of AI agents calls for reaffirmation of responsible practices around AI and stakeholders must work together to create the clear, consistent guardrails needed for the public to trust and adopt AI agents.

To achieve the promise of agentic AI while managing these transitions, technologists and policymakers should embrace a bold and responsible approach, accelerating innovation and adoption to harness agentic AI's immense potential, while strengthening governance principles to help address emerging agentic risks.

This paper explores the **privacy, safety, security, digital commerce, and labor implications** of AI agents, with the goal of giving policymakers a useful framework for understanding and evaluating agentic AI. It looks at implications for both individual users and the broader digital ecosystem. To help policymakers think about effective agentic AI governance, we introduce the **TRUST framework**, with five core pillars of Transparency, Reliability, Unified protocols and standards, Safety and security, and Task augmentation:



Google takes a bold and responsible approach to agentic AI: innovating to harness the technology's immense potential, while mitigating risks through fit-for-purpose frameworks. We encourage policymakers to take a similar approach — supporting the remarkable benefits of AI agents, while evolving regulations to respond to specific problems as they arise.

02 The Evolution of AI: From Answering to Acting

The Evolution of AI: From Answering to Acting

We are stepping into the next fundamental chapter of AI. For the past several years, the public's understanding of AI has been largely shaped by chatbots—generative models that synthesize information, answer questions, and create content. Agentic AI represents a decisive move beyond conversational interfaces into execution engines.

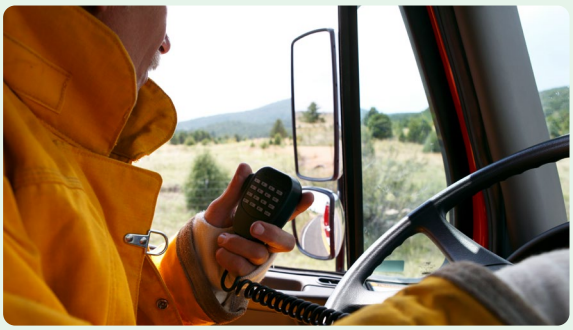
We define AI agents more broadly, as systems that combine the intelligence of advanced AI models with access to digital tools, empowering the agents to take actions on behalf of users, under their control. Throughout this paper, we use 'AI Agents' and 'Agentic AI' interchangeably.

At the core of agentic AI are new capabilities including planning, tool use, and task execution. Agents can use the information they gather to work through a problem logically (i.e., reason); and unlike prior AI applications that only generated content, they can formulate a multi-step plan (i.e., plan), interact with other systems, tools, and websites (i.e., use tools), react to changes in their environment, and autonomously execute complex tasks with appropriate human supervision.

Agentic AI is not a binary switch between purely responsive models and fully proactive systems; rather, agents' autonomy, reasoning and planning capabilities, and goal complexity exist on a continuum. This spectrum spans from assistants that research and curate information for human decision making, to collaborative partners acting under users' close supervision, to advanced agents empowered to independently take actions at users' direction.

By interacting directly with other systems, applications, and websites through APIs and user interfaces, agents expand AI beyond a tool of *inquiry and creativity* to a *tool of execution*. This evolution shifts human-computer interaction from direct management to adaptive orchestration, transforming software from an application you consult into a digital collaborator to which you delegate.

AI Agents in Practice



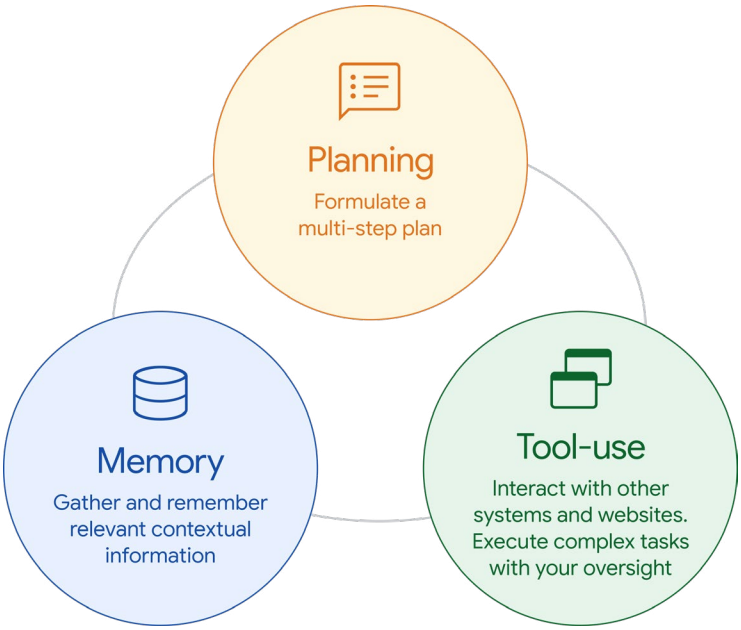
Crisis Response

EvacSafely uses AI and geospatial precision to help public safety agencies navigate crises and protect communities

As climate-driven disasters intensify, many public safety agencies still rely on outdated tools for emergency planning. EvacSafely is using Google's AI tools to help first responders and communities build safety plans. Bringing together advanced geospatial analytics, historical hazard data, and AI-assisted decision support, they pioneered an Evacuation Planning Agent that creates complex, multi-step plans. Under officials' supervision, this agent collaborates in real-time with external sources—autonomously requesting traffic and shelter data—to coordinate a comprehensive response. The technology reduces planning time from 60 hours to 2 hours, allowing officials to run more scenarios and providing residents with clearer, map-based guidance.

What is an AI agent?

AI agents are systems that combine the intelligence of advanced AI models with access to digital tools, to take actions on behalf of users, under their control.



Agentic AI is not a binary switch - it exists on a spectrum

	Less agentic	More agentic
Goal: What does it do?	Complete small, clearly defined tasks	Work toward bigger goals and manage multi-step projects
Environment: Where does it work?	Operate in structured, predictable settings	Navigate complex, real-world situations that constantly change
Planning: How does it plan?	Follow step-by-step instructions	Create plans and adapt as the situation unfolds
Action: What can it do?	Provide information for people to act on	Take action directly while under human control and guidance

03 The Opportunity: Unprecedented Productivity and Value

The Opportunity: Unprecedented Productivity and Value

For individual users, agents promise a productivity windfall, reclaiming countless hours lost to bureaucratic navigation, scheduling, and digital administrative chores. For the broader economy, agents are poised to accelerate commerce, streamline supply chains, and supercharge worker productivity.

Recent economic projections estimate that if organizations adapt their workflows to human-agent collaboration, AI agents could unlock up to $2.9 trillion in economic value in the United States alone by 2030, with agentic commerce alone potentially generating $1 trillion in the US and $3 to $5 trillion globally. We are already seeing this impact: from energy grid operators using agents to drastically speed up the interconnection process of renewable resources, to scientists forming new scientific hypotheses and research plans, to healthcare providers deploying agents to assist with clinical reasoning and patient communication. Agentic AI is no longer a theoretical concept; it is an active economic engine providing tangible societal benefits.

> **AI agents can unlock up to $2.9 trillion in economic value in the United States alone by 2030. Global annual revenue from agentic commerce alone may reach $3 to $5 trillion.**

For organizations across every sector, adopting agentic AI will soon shift from a competitive advantage to a core necessity. In a landscape where teams' velocity increasingly depends on the speed of machine execution, organizations that fail to integrate agentic workflows risk falling behind. Keeping pace means reimagining business processes to leverage digital partners that can operate continuously, adapt dynamically, and manage complexity at scale.

AI Agents in Practice

Public Sector

The Advisory Council on Historic Preservation stewards history while building for the future with Google's AI agents

Modernizing infrastructure while protecting historic properties under the law is a delicate, high-stakes balancing act that typically requires months of manual review. The Advisory Council on Historic Preservation, an independent US federal agency, is streamlining this process using Google AI and geospatial reasoning agents that act as a research and regulatory partner. A multitude of agents retrieve cultural and geographic records across hundreds of federal and state databases, while AI world models generate simulations of how a project might affect cultural resources in the real world. By simulating construction impacts in real-time, these AI systems allow planners to adjust project locations—preserving the past while building for the future.

AI Agents in Practice

Scientific Breakthroughs

Accelerating scientific breakthroughs with AI agents

Co-Scientist is a multi-agent AI system that helps scientists generate novel hypotheses and research proposals. Experts at Imperial College London and the Fleming Initiative, in partnership with Google, used this tool to independently recreate a discovery about how superbugs spread immunity to antibiotics, leveraging decades of data. Ultimately, Co-Scientist could help accelerate the development of life-saving treatments.

04 The Impact Landscape

The Impact Landscape

To navigate these complexities, policymakers and technology stakeholders should focus on agentic systems' network of connections. The impacts of agentic AI can be understood across two primary dimensions: the implications for individual users, and the implications for broader digital and economic ecosystems.

Implications for Individual Users

Delegating tasks with confidence

As users delegate increasingly complex and consequential tasks to AI, reliability becomes a foundational digital safety guardrail for agentic AI. If an agent makes a mistake, the result could be an incorrect bank transfer, a misdirected shipment, or a canceled medical appointment.

Users should have confidence that their agents act within defined boundaries and that their agents will faithfully follow their direction, pause when it encounters ambiguity, and reliably defer to human judgment for high-stakes decisions. Reliability builds trust and trust is the foundation for widespread adoption.

The delegation of authority to an AI agent also carries the risk of unintended actions. A misaligned agent could accidentally delete vital cloud infrastructure, make unauthorized corporate purchases, or send sensitive communications prematurely. To mitigate these risks, an agent's actions should mirror the user's explicit and implicit instructions, combined with oversight mechanisms and user-defined boundaries that prevent an agent from taking unintended, and potentially irreversible, actions. For example, Google has developed tools like the User Alignment Critic to increase an agent's reliability. The User Alignment Critic runs after planning is complete to double-check each proposed action and determine whether they serve the user's stated goal.

As agentic AI matures and becomes more reliable, the opportunity cost of human intervention becomes more pronounced. Many industries are moving from relying on constant human intervention to human supervision. For example, rules for original unmanned aerial systems (drones) required a pilot to physically see the aircraft at all times; however, research demonstrated that Detect-and-Avoid technology became far better than humans at identifying and avoiding hazards, without suffering from fatigue or distraction. This prompted policymakers to move towards human-*on*-the-loop rather than human-*in*-the-loop requirements. AI governance could follow a similar path as agent performance improves, with high-level oversight of an agent's autonomous execution replacing prescriptive, manual approval of every step of an agent's actions. Just as a Waymo car senses and avoids conflicting traffic better than the average human driver, an agent possessing safety guardrails to "detect and avoid" digital hazards—such as policy violations or security threats—can provide improved levels of safety and security.

Learning From Analogs

This is not the first time industries have moved from manual control to automated, scaled execution. Advances in industrial equipment, aviation, commercial drones, and autonomous vehicles provide a roadmap for agentic AI governance.

1. **Applying Existing Regulations**

   Aviation and Occupational Safety and Health Administration (OSHA) requirements use evidence-based frameworks to determine appropriate safeguards. In transportation, existing safety requirements are designed to protect drivers, passengers, and public safety. Municipalities have concluded that road safety rules should continue to apply to all vehicles operating on public streets with or without the presence of a human driver. Agents are no different: If an action is unsafe for a human to perform, it remains unsafe for an agent to perform on a human's behalf.

2. **Operational Design Domains (ODDs)**

   In the context of Autonomous Vehicles, an ODD defines the specific conditions under which an autonomous system is designed to function safely (e.g., specific speeds, weather conditions, or geofenced areas). Agentic actions should be similarly tethered to defined digital ODDs. For example, an agent might be authorized to read and draft emails, but the deployer may strictly geofence it from making automated payments or deleting system files.

3. **From Human-in-the-Loop to Human-on-the-Loop**

   Early drone regulations mandated Visual Line of Sight (VLOS), requiring a pilot to physically observe the aircraft at all times. However, research has proved that Detect-And-Avoid (DAA) technology identifies hazards significantly faster and better than humans do. Consequently, regulations evolved to permit Beyond Visual Line of Sight (BVLOS) operations, acknowledging that safety is better served by DAA technology than by human eyesight. As AI technology matures, high-level oversight may provide more assurance than prescriptive, manual approval.

Safeguarding personal data

To be truly useful, a general AI agent must operate with sufficient context. Personalization transforms agentic AI from a generic tool into a relevant personal assistant, but doing that well requires balancing usefulness with privacy considerations. Depending on the task at hand, access to a user's emails, calendar, chat history, financial preferences, health data, or personal files may be needed for a personal assistant to be helpful. Clear, intuitive, and user-centric transparency and control will remain critical, as they have been for the previous generation of consumer technology. However, increased agentic AI use introduces more complexity into the data chain of custody. Attempting to provide granular control, for example requiring user consent at most stages, will compound the phenomenon of "consent fatigue" or "consent blindness" long observed in the online ecosystem, with users mindlessly clicking "accept" on endless pop-ups in order to access a service. Meaningful privacy in agentic systems should instead be based on overarching, contextual transparency and controls that honor users' intent, preferences, and reasonable expectations of privacy in any given context, while preserving the seamless experience that defines helpful assistants.

The need for and scope of data access will also depend on an agent's scope. There is a distinction between general agents designed to be broad personal assistants across a wide range of tasks, and specialized agents (e.g., coding agents) that are typically good at a narrow range of tasks. A general agent might need access to a broader range of data compared to a specialized agent, but both should offer robust data protection mechanisms.

Evolving user oversight in this way will require technology that can be trusted to execute on users' evolving expectations of privacy. If users are no longer able to oversee every single instance of data sharing as an agent executes a task on their behalf, they must be confident that technical protections will apply and enforce their privacy preferences at the right junctures. The past two decades have brought considerable progress in the field of privacy-enhancing technologies (PETs) in terms of development and deployment. Examples include Trusted Execution Environments (TEEs) and Differential Privacy, which have been integrated into a wide range of AI products. Continued investment in this field, and government incentives for adoption of these techniques, is critical.

Agents also open up new areas for exploration. For example, we could imagine that agents communicate privacy preferences or data sharing permissions to one another; we could also imagine a user privacy agent tasked specifically with enforcing a user's stated privacy preferences. Moving towards more context-based, expectations-based privacy protection within seamless agentic experiences will require the technology itself to shoulder more of the burden for demonstrable data protection.

Implications for the Ecosystem

Keeping the Web Running, Securely

The internet was built for human navigation, characterized by clicks, page loads, and visual interfaces. Everything from user interface and experience designs to backend protocols and cybersecurity practices were built around the speed and patterns of human users. Agents however, operate at greater speed and scale, requiring updating interaction protocols and agent-compatible web and mobile standards to keep the web running smoothly. Additionally, agents interacting with each other and with many tools and helping on a large set of tasks means that threats and risks have to be managed at scale.

Threats will scale with agent capability and deployment context. More capable AI agents inherently pose larger risks, as their superior reasoning allows them stronger abilities to potentially override security measures and execute complex strategies if they become compromised. Managing this risk requires scaling oversight and monitoring. And as deployments scale into multi-agent environments rather than isolated silos, agents built on AI models with varying degrees of safety and security will inevitably interact. Industry will need to build toward common, ecosystem-wide security standards and protocols to protect such heterogeneous systems.

The tech sector has successfully addressed the need for common protocols countless times before. Real-time voice, video, and peer-to-peer data sharing over the web was enabled by the WebRTC protocols; and the transition to HTTPS encryption effectively has made the web secure-by-default for billions of users. In the agentic era, the tech community is making similar progress. Google has launched Agent2Agent (A2A), an open protocol facilitating collaboration in a dynamic, multi-agent ecosystem across siloed data systems and applications, and the Agent Payments Protocol[1], an open protocol designed to help securely initiate and transact agent-led payments across platforms, and we are actively adopting standards developed by others in the community, like the Model Context Protocol and Agent Skills. Recognizing the need for industry-wide, updated protocols and standards, the U.S. National Institute of Standards and Technology (NIST) has launched the AI Agent Standards Initiative to ensure AI agents are secure, interoperable, and trusted for widespread adoption.

Google's Secure AI Framework (SAIF) divides AI security risks into four component areas: Data, Infrastructure, Model, and Application. While some approaches to AI security focus primarily on the model, SAIF addresses risks and controls throughout the entire AI development lifecycle. An agent's level of autonomy governs the potential severity of a security failure; the more an agent can dynamically orchestrate multi-step actions without guardrails, the greater the risk of failure. For AI agents, SAIF has developed a specialized diagram that examines unique risks that agents introduce.

Agentic AI enhances both cyber offense and cyber defense. On the offensive side, bad actors can deploy malicious agents capable of executing sophisticated, automated cyber-attacks at a speed that vastly outpaces human defenders. As observed with Claude Mythos, AI models can excel at vulnerability discovery, even without being purpose-built for the task. Automated vulnerability discovery and remediation capabilities will be eventually integrated into the software development life cycle, and code will be more difficult to exploit than ever; however, until this transition happens, threat actors may be able to weaponize this capability to discover and exploit novel vulnerabilities. Within this transition window, defenders have two critical tasks: leveraging AI to harden current software as rapidly as possible, and preparing to defend systems that have not yet been hardened. Now is the time to strengthen incident response playbooks, reduce exposure, and incorporate AI into security programs.

To this end, Google offers a wide range of AI-integrated defensive ecosystem solutions. Security agents such as Big Sleep or CodeMender can detect and patch vulnerabilities far faster than software engineers alone. Google Threat Intelligence delivers detailed and timely threat intelligence to security teams around the world. Mandiant Security Consulting Services helps organizations design and operationalize cybersecurity and threat intelligence architecture. Agentic SecOps provides the foundation for an agentic security operations center. Mandiant Threat Defense leverages frontline intelligence and AI-enabled telemetry to proactively identify and disrupt advanced, machine-speed threats. Google Cloud Model Armor helps customers proactively screen inputs and outputs to help block prompt injections and sensitive data leaks to secure the AI agents organizations deploy.

The recent popularity and wide-spread adoption of the open-source agentic framework, OpenClaw, demonstrate the appetite for people to use agentic systems in their everyday lives while concurrently highlighting the dangers individual users and the wider agentic ecosystem are exposed to when adoption happens without adequate security awareness, preparedness, and guardrails. That is why agentic security principles, such as those laid out in the CoSAI Principles for Secure-by-Design Agentic Systems, which also guide Google's internal implementations, become even more relevant for individual users and industry as a whole. To secure agentic systems, Google applies a hybrid defense-in-depth strategy that combines deterministic security

1 Google is donating Agent Payments Protocol to the FIDO Alliance to support the future of secure, agentic payments.

measures with dynamic, reasoning-based defenses, providing multiple layers of protection.

And autonomy is not synonymous with anonymity. For example, in the Agent2Agent protocol, agents carry Agent Cards, which contain an agent's name, description, capabilities, provider information, and credential information. Agent Cards act as "license plates for agents" ensuring that agentic AI autonomy does not become synonymous with anonymity. Identifying and labeling agentic activities, integrating with different existing agent identity protocols (such as SPIFFE and Web Bot Auth), and building flexible controls to challenge and block agents based on identities and behaviors is an active area of research. Interoperable agent identifiers, with cryptographically robust authentication methods, could enable fraud prevention, allow tracing of security incidents, and help distinguish trustworthy agents from malicious bots or compromised agents. Many industry actors already implement agent identification solutions on their platforms; for example, Gemini Enterprise Agent Platform lets developers assign unique identities to their agents to authenticate and authorize agent actions to enhance security. As the interactions between websites, agents, and browsers evolve, agent identification and authentication solutions may need to adapt to keep the web both open and secure.

To keep pace with the speed and scale of AI-enabled attacks from threat actors, collective and coordinated action across companies and borders will help give cyber defenders agentic tools that will give them the advantage. Industry coalitions, such as the Coalition for Secure AI (CoSAI), play an important role in shaping best practices and strengthening the security of AI and agentic systems across the broader ecosystem.

AI Agents in Practice

Cybersecurity

Helping find security vulnerabilities before they can be exploited

Cybersecurity has long been an uneven battle, with defenders struggling to protect massive digital landscapes while attackers only need to find a single crack to succeed. To flip this script, Google's Big Sleep AI agent acts as a proactive defense tool that finds hidden software flaws before they can be exploited. In a 2025 breakthrough, Big Sleep identified a critical vulnerability that hackers were already planning to use—the first time AI foiled an attack before it started. By instantly analyzing code that would take humans weeks to review, Big Sleep replaces slow manual searches with rapid, automated defense, securing the broader internet and stopping threats in their tracks.

Transforming Digital Commerce

Agentic commerce marks a shift from manual, user navigation involving searching, comparing, and authenticating across platforms, to agents that autonomously analyze merchants' offerings at speed and scale, under a user's direction. In an agentic future, personal AI assistants will likely communicate directly with businesses' agents to execute complex, multi-step transactions. Consider a routine task like restocking a home pantry. Instead of a user manually browsing various grocery apps for the best prices and delivery slots, their personal agent could autonomously negotiate with multiple merchant agents, balancing real-time inventory, bulk discounts, and the user's dietary preferences to finalize a purchase that optimizes both cost and convenience, while saving the user time.

While agents can help deliver a frictionless customer experience, they can also enable new abuse and fraud vectors. Consider a high-demand product launch: 10,000 individual customers task their personal AI agents to each buy one item the moment it drops. This is great for a merchant reaching thousands of new customers. But consider the risk of one malicious scalper deploying 10,000 agents to buy the entire inventory for resale. To a traditional security system, both scenarios look like an identical "attack." On the other hand, traditional point-of-sale fraud detection signals can become obsolete when the buyer is an agent. To respond to these challenges, fraud and abuse prevention mechanisms must evolve to stay alert to agents' activities. For example, Forensic Intelligence provides specialized protection against malicious automation, account takeovers, and payment fraud. Google has also developed a technical framework for agents identity and security, enabling a safe agentic web that protects users and businesses from fraud and abuse while helping deliver an autonomous and frictionless agentic commerce experience.

This paradigm shift also unlocks massive opportunities. It's possible to imagine a future where if a user chooses to share their unique preferences with their agent; over time their agent will learn when to flag deviations, like a merchant agent attempting a stealthy price-bait-and-switch that the user might have otherwise overlooked.

Trust is a two-way street. Traditional commerce relies on Know Your Customer principles. Merchants should be able to identify malicious activities, regardless of whether those activities are mediated by an agent or not. Conversely, consumers must be able to verify the identity and legitimacy of the merchant or advertiser to avoid a spoofed storefront. This creates a "circle of trust" for the entire transaction loop. To facilitate merchant and advertiser verification, replacing uploads of physical IDs and business documents with deterministic cryptographic credentials will enable platforms to verify legitimacy in seconds with high assurance, surfacing that information to consumers (and the agents operating on their behalf, all while under their control) and making for a more secure and accountable ecosystem.

The future of commerce relies on open standards that support secure, platform-agnostic machine-to-machine transactions. The Universal Commerce Protocol (UCP), for example, is a new protocol that is open and platform-agnostic, meaning it can be deployed on any platform to support users' entire shopping experience. It is a standard independent of any single company, agent, or payment provider and is designed to make AI-powered shopping work better for consumers, retailers, and marketplace or intermediary platforms; to this end, UCP establishes a common language and building blocks for AI shopping agents, online stores, and payment providers to communicate, eliminating the need for unique connections for every individual agent and it can be tailored to any platform or service.

Navigating Workforce Transitions

The rapid advance of agents has raised concerns about AI's impact on the future of work. This shift is visible in software engineering, where the technology is already fundamentally reshaping how code is written, and thereby changing the daily work of many engineers, while expanding the universe of people able to "vibe code".

Jobs include both bundles of tasks and a coordination layer, coordinating those tasks to meet dynamic social and economic demands. Although many digital tasks may be automatable, the coordination layer includes navigating interpersonal relationships; acting on shared, tacit knowledge; and exercising judgment about unique circumstances. Agents are not people, and will struggle to replicate the way a person brings many pieces together to achieve goals, including the judgment to know when goals need to change. As the cost of agentic task execution drops, the economic value of knowing which tasks to automate and why and the ability to coordinate across tasks is likely to increase.

Variation in which tasks agents can replicate helps explain why agentic AI has advanced furthest in software engineering. Coding exists within a highly structured, logical framework defined by immediate feedback loops and "right" answers. In contrast, transitioning agents into non-coding roles—such as management, creative strategy, or complex operations—will likely prove more difficult. These fields are inherently "unstructured," requiring humans to navigate ambiguous data and the subtle interpersonal dynamics that algorithms struggle to solve.

History — from the Industrial Revolution to telecommunications to computers to the internet — suggests that transformative technologies ultimately expand the labor frontier rather than shrink it. The AI transformation is already creating new professional categories: LinkedIn reports that employers have created at least 1.3 million AI-related job opportunities in the past two years, including data annotators, AI engineers, and forward-deployed engineers.

At the same time, agentic AI will drive change in many sectors of the economy, albeit unevenly. Navigating this shift requires a new partnership across government, industry, and civil society to encourage AI adoption that augments human capability rather than replacing it. Technology firms and policymakers can work together on scaling skills training, assessing AI's economic impacts, and gathering data-driven insights to ensure every worker benefits from the transition. For example, Google's AI Opportunity Fund supports organizations training 720,000 workers across Asia-Pacific and 100,000 Micro-, Small-, and Medium-sized Enterprises across Southeast Asia to accelerate AI readiness across the region.

Google is also actively investing in the American economic backbone by providing specialized AI training to educators, small businesses, and thousands of workers in manufacturing and skilled trades. This includes a $10 million grant to the Manufacturing Institute to integrate AI into vocational training and train 40,000 US manufacturing workers, No-cost AI Professional Certificates and specialized training for small businesses, and AI literacy training, which provides free of charge, comprehensive Gemini training to all 6 million K-12 and higher education faculty in the U.S. Coupled with initiatives like the AI & Economy Research Program to offer data-driven insights for navigating the evolving nature of work, these collaborative efforts lay the foundation for smart policy approaches that empower the workforce and drive widespread economic growth.

05 Guiding Principles for Governance of AI Agents: The TRUST Framework

Guiding Principles for Governance of Agents: The TRUST Framework

Agentic AI is a rapidly evolving technology. Because we are at the beginning of agentic AI development and deployment, any policy framework must remain sufficiently agile to ensure that regulatory guardrails do not inadvertently stifle the innovation driving technological breakthroughs.

To go fast in the long run, regulators must first go slow, using a methodical, evidence-based approach rather than a rush to regulate that often leads to institutional friction and premature complexity. Early issues with the EU framework show that balancing speed with clarity in AI regulation is a significant challenge, sometimes leading to unintended consequences and unnecessary complexity for emerging technology. While existing legal frameworks can provide a foundation, they should be applied appropriately, with an eye for second-order effects and trade-offs between concerns like competition, consumer protection, privacy, and content moderation.

We already have robust frameworks in many of these areas, and agents aren't exempt — if an action is illegal without an agent, it is illegal with an agent. By recognizing that the focus should be on regulating the *actions* of agents rather than micromanaging *how* they operate, policymakers can build a harmonized, predictable policy environment for the technology to evolve. Using pilot programs and regulatory sandboxes, policymakers can gather data on agentic behavior, system failures, and economic impacts to guide evidence-based policy, which is the best way to navigate Collingridge's Dilemma — that it's possible to regulate new technology too soon as well as too late.

To help policymakers and industry leaders navigate the emerging agentic AI landscape, we propose the TRUST Framework—a blueprint for governing AI agents, structured around five core pillars: Transparency, Reliability, Unified Protocols, Standards, and Task Augmentation.[2]

AI Agents in Practice

Public Sector

Regulatory sandboxes help governments test AI agents in a highly-controlled environment.

The Singapore Government partnered with Google to create a global-first sandbox to identify and test agentic solutions for public sector use cases in a safe, simulated environment. A key use case focused on how these tools could help citizens apply for social assistance, a process that can often feel complex and overwhelming. By acting like a digital assistant under users' supervision, the prototype could read informal notes, fill out complex forms, and even calculate monthly income from various sources. This promising partnership proves how AI agents can simplify administrative hurdles and improve public service delivery, helping everyone access the support they need.

2 We note that researchers have previously proposed a distinct TRUST-AI framework for healthcare settings, which focuses on responsibly evaluating AI systems within clinical settings. In comparison, the TRUST framework in this whitepaper concerns the governance of AI agents.

Trust Pillar: Transparency
Theme: User Notice, Observability & Adaptable Oversight

Users should be able to understand the actions their agents take on their behalf, through clear notices, appropriate controls, and context-appropriate human oversight. Agents' access to data should depend on the scope of what users need them to do.

Users should have adaptable, context-appropriate oversight on agents' key actions and decisions. This enables users to intervene and course-correct agents' actions as needed. Developers should offer a range of user options where appropriate, from requiring explicit human intervention for high-stakes actions, to general supervision, to post-action verification. While human-agent interactions should shift from manual "human-in-the-loop" to high-level "human-on-the-loop" as agent reliability scales, explicit human intervention should continue to be an option for high-stakes actions carrying significant risks.

Observability—verifying that the agent correctly followed instructions and achieved the intended outcome—and best practices for human intervention are currently active areas of work across the industry. Google's agentic safeguards and controls seek to give users oversight of agents' key actions and decisions. As we expand our agentic features, we are exploring ways to minimize friction in the user experience by allowing users to set adaptable boundaries, such as defining financial limits for routine tasks or requiring confirmation for specific categories of action. This flexibility allows users to explicitly authorize and automate agentic actions in line with their needs and risk tolerance, while ensuring the agent operates within the bounds of system-level safety guardrails.

Additionally, stakeholders should continue to invest in privacy-enhancing technologies, such as hardware-based Trusted Execution Environments (TEEs) and Privacy Preserving Measurements, and innovate with agentic technologies that can help enforce data protection rules and preferences.

Policymakers should avoid blanket oversight mandates, and instead recognize that user intervention should scale with the potential impact of the agent's actions.

Trust Pillar: Reliability
Theme: Staying within defined boundaries & following user's direction

Users should have confidence that their agents follow their direction and act within defined boundaries. Restrictions on the range of actions an agent can take help prevent them from taking unintended, harmful, or policy-violating actions.

Users should help determine and tailor the boundaries of agentic actions; for example, allowing an agent to access information about a user's allergies helps the agent curate a grocery list that is safe and more helpful. User expectations for acceptable ranges of action will likely evolve over time, and based on agents' reliability and capability levels and the context of specific application (e.g., low vs. high impact). Consequently, the types of tasks an agent is allowed to perform should remain flexible, letting users tailor the limits of their agents' actions.

Because of the sheer scale and speed of AI agents' operations, traditional mitigations that rely on human oversight and intervention could quickly become too costly to maintain. Consequently, defenses must evolve beyond traditional, human-driven mitigations toward automated, scalable safeguards.

Policymakers should leverage tools like regulatory sandboxes and pilot programs for high-speed, high-stakes agent deployments (e.g., finance, energy grids) to collect real-world data on failure modes, and encourage public-private incident reporting of system failures, for example, by providing the appropriate infrastructure for voluntary reporting, in order to inform adaptive regulation.

Trust Pillar: Unified Protocols
Theme: Interoperability & Open Standards

Agents work best in a collaborative ecosystem with common standards. As the protocols underpinning the internet and wi-fi demonstrate, common technical standards for information exchange and cooperative action reduce friction and create exponential value. Similarly, agents will benefit from global cooperation on consensus protocols, which will allow them to securely connect to various data sources, services, and platforms, and to collaborate with each other.

An open ecosystem requires a foundation of trust. Common standards would reduce the potential for abusive agent behavior, such as automated astroturfing (deceptive use of bots to fake public opinion) or personalized spam. The industry must advance common, open, and secure protocols to ensure agents from different providers can work with each other seamlessly and securely. Technical standards (like those being developed through bodies like IETF) could help website and app owners distinguish between trustworthy agents (like those booking travel or ordering food) and malicious bots and set appropriate boundaries on agents' access.

Governments can prevent a fragmented regulatory environment by championing open technical standards for agents. They can drive the market toward trusted, verifiable agents by promoting standards on secure agent interactions as well as agentic commerce standards to cultivate trust in an open ecosystem and ensure technology adoption is beneficial, reliable, safe, and secure for users and businesses alike. Governments can encourage market-wide alignment by referencing industry-led, consensus-based standards in government policy frameworks and public procurement guidelines.

Trust Pillar: Safety and Security
Theme: Security by Design & Benchmarking for Safety

Industry principles and standards that emphasize a secure-by-design approach can help create a consistent baseline for security against efforts by bad actors trying to subvert or redirect agents' activities. Making security a core part of procurement will encourage vendors to build robust security features into their commercial products and raise the security baseline across the market. In parallel, safeguarding user privacy is essential for building and maintaining public trust. Privacy protections must rely on contextual transparency and controls, meaning data permissions should scale with the complexity and context of the agent's task.

As with other technologies, Google takes a gradual and responsible approach to deploying agentic AI, assessing risks and designing risk-calibrated testing strategies to identify and mitigate risks. This may include: conducting research on several prototypes, working with trusted testers and external experts to develop and apply best practices, performing risk assessments, safety evaluations, and red-teaming to stress-test systems before they reach users. To better identify and mitigate real-world harms, we may use human-in-the-loop live testing, scaled safety evaluations, and sandbox environments, using the most appropriate testing methods for each product. Sandbox environments replicate complex, digital experiences and adversarial attacks, allowing us to identify and mitigate harmful risks without exposing the public web. In our red-teaming efforts, creative experts actively attempt to circumvent system safeguards to uncover unforeseen vulnerabilities. This process is augmented by our Vulnerability Reward Program that invites security researchers to further discover and report AI-related vulnerabilities. Additionally, our scaled safety evaluations subject the system to thousands of test cases, measuring its reliability across the full range of situations it may encounter. While it's impossible to identify all the risks, this work helps us address novel safety issues before they materialize.

The technical community is developing a range of evaluation benchmarks and frameworks for assessing agents' capabilities, reliability, safety, and security. These efforts, however, should further expand to span highly

Trust Pillar: Safety and Security (continued)
Theme: Security by Design & Benchmarking for Safety

specific, industry-tailored environments, for example, assessing the distinct risks of healthcare agents or finance agents.

Governments can also help develop comprehensive agentic evaluation benchmarks and testing methods by (a) expanding benchmarks in high-risk application areas (where the private sector does not have enough data to build benchmarks), (b) working with other governments towards creation of common, standardized benchmarks for agents, and (c) building platforms for safety-relevant evaluations, including tests for cyber-capabilities and chemical, biological, radiological, and nuclear (CBRN) risks, without exposing the underlying sensitive data. For example, the UK AI Security Institute has published Inspect, a framework designed for evaluations of large language models and agents scaffolds. This open-source evaluation framework enables rigorous, large-scale testing of AI agents across diverse and complex tasks, leading to advances in agentic safety and security.

Additionally, future multi-agent systems, where agents work with each other, may introduce emergent failure modes such as miscoordination, conflict, and collusion that cannot be predicted by evaluating an AI agent in isolation. Likewise, governance of risks from multi-agent dynamics would benefit from better evaluations and benchmarks.

These efforts should prioritize the creation of common, standardized benchmarks, ensuring that governments work toward shared goals rather than establishing a fragmented landscape of duplicative, individual mandates for evaluation.

Requiring agent evaluation details as a part of the procurement process would encourage vendors to prioritize rigorous safety testing and adopt standardized benchmarks as a minimum requirement in their development process.

Trust Pillar: Task Augmentation
Theme: Human Augmentation, Upskilling & Workforce Transitions

By preparing workers to manage AI, public and private sector actors can mitigate labor disruption and maximize the productivity benefits of human-AI collaboration.

Agentic AI will drive change in many sectors of the economy, albeit unevenly. Navigating this shift requires a new partnership across government, industry, and civil society to encourage AI adoption that augments human capability rather than replacing it. Technology firms and policymakers can work together on scaling skills training, assessing AI's economic impacts, and gathering data-driven insights to ensure every worker benefits from the transition.

Guiding Policy Levers

Governance

- ☐ Rely on existing frameworks such as consumer protection and applicable sectoral laws, such as in finance and health, recognizing that what is illegal to do without an agent, is still illegal to do with an agent.
- ☐ Adopt an outcome-focused approach, focusing on the results of AI actions rather than the underlying technology.
- ☐ Conduct cross-agency regulatory stocktaking and leverage stakeholder consultations (e.g., requests for information) to identify specific legal gaps or barriers to agentic adoption.
- ☐ Prioritize going slow at the beginning to go fast in the long run: apply a methodical approach to oversight to prevent the "rush to regulate" that often leads to institutional friction and premature complexity.

Transparency & Reliability

- ☐ Establish controlled environments (i.e., regulatory sandboxes) for high-stakes agent deployments (e.g., finance or energy) to collect real-world data on failure modes and effective guardrails.
- ☐ Encourage incident reporting and build the necessary infrastructure for voluntary, public-private incident reporting of agentic systems failures to inform adaptive regulation.
- ☐ Avoid blanket oversight mandates; instead, ensure that the level of human intervention (e.g., human-in-the-loop vs. human-on-the-loop) scales with the potential impact of the agent's actions.

Unified Protocols & Standards

- ☐ Promote open technical standards, developed to keep the web open but secure in the agentic era.
- ☐ Reference industry-led, consensus-based standards in government policy and public procurement guidelines to encourage market-wide alignment and interoperability.

Safety & Security

- ☐ Work with other governments towards creation of common, standardized benchmarks for agents, including cyber- and CBRNE risks and capabilities.
- ☐ Build platforms for agent evaluations that enable rigorous, inspectable, large-scale testing of agents across diverse and complex tasks.

Task Augmentation & Workforce Transition

- ☐ Partner with industry and civil society to fund and scale AI literacy and vocational training programs for workers and educators.
- ☐ Gather data-driven insights, for example, by commissioning research into identifying the sectors which require the most support during the workforce transition.

06 Conclusion: Go Slow to Go Fast

Conclusion: Go Slow to Go Fast

If we navigate this transition successfully, the ultimate promise of agentic AI extends far beyond automating calendar invitations or accelerating digital checkout lines. By pairing human ingenuity with the tireless execution capacity of agentic AI, we can reimagine energy grid resilience, revolutionize personalized healthcare, and leverage mass logistics to respond to natural disasters, and much more.

As profound as the technology transitions could be, agentic AI does not necessitate an entirely new legal frontier, nor do its challenges require policymakers to reinvent regulatory goals or fundamental frameworks. Current regulatory principles around privacy, consumer protection, and digital security still apply: if an action is unlawful to do without agents, it is unlawful to do with agents. Because agentic AI could blur the traditional boundaries between regulatory domains, inter-agency collaboration (e.g., between commerce, tech, and national security agencies) is essential to avoid fragmented views and ensure a cohesive and effective approach. The unique nature of delegation—where an AI acts autonomously or semi-autonomously—means that agentic AI adoption would benefit greatly from regulatory stocktaking, assessing applicability, and updating existing legal frameworks if gaps are identified. Policy must evolve to address the scale, speed, and autonomy of these systems.

Before moving to create entirely new regulatory frameworks, industry and policymakers should study the ways agents interact with users and broader ecosystems. Since agentic AI is a nascent and rapidly evolving technology, we encourage policymakers to:

- Rely on existing legal frameworks, such as privacy and applicable sectoral regulations.
- Assess the sufficiency and applicability of existing laws, such as consumer protection regulations, and update them if gaps are identified.
- Ensure legal frameworks are outcome-focused and technology-neutral.
- Leverage tools like regulatory sandboxes to gather real-world insights about agentic AI.
- Allow developers and deployers the flexibility to continue research and development (R&D) and innovation.

Understanding where agentic AI is and is not different from past transformative technologies can allow for adaptation as needed.

The transition to agentic AI is not a distant possibility; it is underway and already opening new digital and economic opportunities. To harness this potential and capture the trillions of dollars in economic value and productivity that agents offer, policymakers, industry leaders, and civil society should focus on how to govern agentic AI by building on existing standards and regulatory frameworks. The TRUST framework will help ensure privacy, security, and human oversight are core guardrails for agents, and help the world build an agentic future that works securely and reliably.