Chilling autonomy: Policy enforcement for human oversight of AI agents

Peter Cihon [1]

Abstract

Existing law and policy consensus governs the development and deployment of AI agents. In correspondence with this policy foundation, and in contrast to the ambiguous concept of agenticness, this paper centers autonomy and its constraint, human oversight, as a conceptual basis for AI agent governance. The paper takes initial steps to articulate a standard of care for AI agent development and deployment that can be strengthened and enforced by multiple stakeholders, including policymakers, researchers, and developers. Existing policy, including the EU AI Act, tort law, consumer protection, and cybercrime law, offers means to chill unsafe autonomy today. The paper develops a research and enforcement agenda, and provides a baseline by which to evaluate novel governance proposals.

1. Introduction

The development and deployment of increasingly autonomous AI systems warrants governance scrutiny. In contrast to dominant applications of foundation models today such as chatbot oracles or task-specific tools, AI agents are AI systems capable of flexibly planning and acting on goals to impact the environment over time (Wang et al., 2024). Adoption of AI agents is a function not only of their real-world performance but also the expectations of users and regulators. This political economy of AI deployment steers capital and research investments for AI agent development.[1] Thus, policymakers and other stakeholders can play a role today in shaping the market for and development of AI agents to ensure that increasingly autonomous AI systems do not pose risks that would be democratically unacceptable.

Clarifying how existing law applies to AI agent development and deployment provides a foundation for governance. As discussed below, this foundation can be further strengthened through interpretive guidance and policy research. The policy context in which AI development is happening today does not recognize AI systems as anything other than tools built and used by responsible humans. Barring concerted legal change, governance of AI agents will reflect the governance of AI (and software) systems more broadly. This current paradigm is a useful one; it centers human oversight and accountability.

Today human oversight is a principle of established global AI policy consensus. Human oversight of AI development and deployment is enshrined in the OECD AI Principles (2019) agreement, endorsed by 47 countries and all members of the G20, and echoed in the Bletchley Declaration (2023) and Seoul Ministerial Statement (2024).[2] Laws globally expect humans to be in the loop for intellectual property protections to apply.[3] Autonomous vehicles are deployed with express permission from regulators (CADMV, 2024; Eastman et al., 2023). Regulation of high-frequency algorithmic trading on stock exchanges requires supervisory processes for deployment and the pause of all trading on an exchange when warranted.[4]

[1]GitHub, San Francisco, USA. Correspondence to: Peter Cihon <petercihon@gmail.com>.

The views expressed in this paper are those of the author and do not reflect the official policy or position of their employer.

Workshop on Generative AI and Law (GenLaw '24) at the 41st International Conference on Machine Learning, Vienna, Austria. 2024.

[1]Law, norms, and market demands, not simply capabilities will shape agent development and deployment (Lessig, 1999).

[2]The notable exception to this consensus lies in lethal autonomous weapons, where efforts to ban such systems internationally continue to face challenges (Wareham, 2020), with researchers calling for prohibitions against deployments without human oversight (Simmons-Edler et al., 2024). AI agents present widespread use and concern domestic regulatory motives in ways that military-specific systems do not. Given these concerns and the global consensus today, a lawful adoption race that undercuts human oversight is not likely. Policymaker vigilance to continue to center human oversight will make it less likely still.

[3]E.g., the U.S. Copyright Office (2022) rejected protection for an autonomously created artwork and the UK did similarly in a patent case.

[4]See, e.g., Securities and Exchange Commission (2020) on the Market Access Rule and Limit Up-Limit Down plan.

Where not expressly mandated or defined in law, human oversight manifests in a standard of care that defines the safe development and use of increasingly autonomous AI systems. The standard of care is norm-based governance with legal teeth: it reflects reasonable actions and precautions, where the failure to adhere to a standard of care that results in harm brings tort or regulatory liability.[5] This reasonableness standard reflects the practices of other actors, and thus may be informed by industry best practices, academic research, policymaker statements, and legal requirements. Notably, the standard of care is a flexible concept that can be strengthened by stakeholders articulating new states of the art. Thus, governance research and policymaker attention have an important role to play in articulating and advancing a standard of care for the reasonable development and deployment of AI agents. This paper begins work to articulate, and centers human oversight within, that standard of care.

As capabilities improve, systems may effectively function with greater autonomy, and this could change inherent needs for and forms of human oversight.[6] Yet, any such changes would need to be negotiated within the political economy of deployment, including user and regulator expectations, which requires human oversight. Thus, not only does human oversight provide a foundation for AI agent governance, it also provides a baseline to evaluate novel proposals from AI developers and governance researchers. Proposals[7] should articulate why changes to human oversight and accountability are needed and why such changes would not pose unacceptable risks.

This paper takes initial steps to define the role that policy stakeholders can play today to shape the market for and development of AI agents. They can do so by articulating the standard of care, offering interpretive guidance for existing laws, and applying scrutiny to increasingly autonomous systems. These steps can steer agent development and deployment in a direction that retains human agency and accountability. This would chill autonomy by encouraging the development of AI agents with affordances that support reasonable human oversight and scrutinizing efforts that deviate from the standard of care and established policy consensus.

The paper proceeds as follows. Section 2 reviews the existing agent governance literature and articulates a tentative standard of care for human oversight. Section 3 identifies ways that the standard of care may be enforced by analyzing the EU AI Act, tort law, consumer protection and cybercrime law. Section 4 concludes with a discussion and research agenda.

This paper makes the following contributions:

- Conceptually, it grounds the governance of AI agents in existing policy consensus of human oversight, and thus articulates a foundation from which to build and baseline by which to evaluate governance proposals.
- Practically, it bridges the gap between theoretical governance and policy implementation. It clarifies existing governance tools, namely the standard of care and policy guidance, that multiple stakeholders can contribute to and use to govern AI agent development and deployment.
- Analytically, it provides policy interpretation to identify several promising means of enforcement today and flexibly over time.

2. A standard of care for human oversight

Current capital and research investments will yield increasingly capable AI agents over time. Today AI agents are largely marketing aspiration, though tool-use and scaffolding software on foundation models have enabled systems of modest capabilities in research settings (Yang et al., 2024; Huang et al., 2023), often neglecting costs (Kapoor et al., 2024).[8] Increasing agent capabilities could take the form of planning over longer time horizons and taking an increasing range of complex actions. More capable AI agents may take actions effectively in increasingly complex and unbounded environments, including the internet at large and physical spaces. As planning, actions, and deployment environments increase in complexity, so too do the risks.

Increasing AI agent deployments and capabilities pose increasing risks, exacerbating current harms and introducing new severe harms. These include reckless use risks, where users deploy AI agents, without adequate information or oversight, only for outputs to harm themselves or others subject to the AI system. They include systemic risks; for example, on the internet, a further increase in human-like bots would further undermine trust in online activity and strain digital infrastructure.[9] Agents capable of completing full chains of tasks would pose challenges and opportunities for the future of work (Eloundou et al., 2023). The risks also include malicious use, the most extreme of which are already influencing policy. The Biden Executive Order on

[5] See, e.g., definition in Wex (2021).

[6] Morris et al. (2023) state that more capable systems enable more autonomy, "though lower levels of autonomy may be desirable for particular tasks and contexts… Carefully considered choices around human-AI interaction are vital to safe and responsible deployment of frontier AI models" (8).

[7] Such ideas have included legal personhood and financial infrastructure to enable AI agents to operate without human oversight.

[8] Widely used chatbot systems today may have access to some tools, but are not directly capable of flexible planning and taking actions in environments.

[9] See e.g., Zittrain (2024).

AI (2023), Hiroshima Process International Code of Conduct (2023), and other documents emphasize risks from loss of "human control or oversight," systems that significantly lower barriers to "design, synthesize, acquire or use" weapons of mass destruction, and systems that automate "vulnerability discovery and exploitation" for cyber attacks. These risks either invoke system autonomy expressly or scale implicitly with increasing automation across multiple actions in a chain of malicious tasks. Related evaluation efforts assess autonomous performance on a series of specific tasks (UK AI Safety Institute, 2024; METR, 2024).

To-date, a small but growing AI governance literature has considered AI agents. Some have focused conceptually on capability-dependent characteristics, emphasizing "agenticness" of such systems.[10] This literature has surfaced worthwhile questions for developers to consider in designing AI agents (Shavit et al., 2023), presented design concepts that could support governance (Chan et al., 2024), and explored the ethics of advanced agents, emphasizing the need for safety investments, including methods of scalable oversight (Gabriel et al., 2024). Another paper looks to the economic theory and law of agency relationships for governance inspiration, taking the perspective that AI agents mark a paradigm shift of AI systems from tools to actors (Kolt, 2024). Still others have called for red-lines in the development of hypothetical agents capable of long-term planning (Cohen et al., 2024).

However, this research would benefit from grounding in prior literature and current policy. Human oversight and the level of autonomy (Simmler & Frischknecht, 2021; Yang et al., 2017) permitted in use are more useful concepts to understand how such systems are and will be governed. These concepts center accountability in the human user and developer, connect to existing literature on humans in the loop and meaningful human control (Crootof et al., 2023; Robbins, 2023; Horowitz & Scharre, 2015), and avoid speculation about systems best left to technical literature. Human oversight entails both information required to responsibly use an AI agent and control it during operation. Oversight stands in opposition to autonomy, and the two can be considered poles on a spectrum: close human oversight constrains autonomy.

Chilling autonomy reduces risks. Human oversight over deployed AI agents maintains existing mechanisms of accountability,[11] permitting current law to address instances of reckless or malicious use.[12] Reduced deployments of AI agents without oversight would similarly reduce systemic risks from their interactions in the environment. Failures will undoubtedly occur, where increasingly capable systems subject to human oversight bring risks of their own, as has been notably seen in human takeovers of semi-autonomous driving. Thus, caution and scrutiny in proceeding towards such developments and deployments would enable societal learning and adaptation as actors and governance institutions process failures as they occur over time. Note that general progress in AI development does not necessarily entail autonomy: increasingly capable AI tools can assist humans with specific tasks or perform general purpose tasks at human direction. A standard of care for human oversight of AI agents can clarify expectations that reduce risks and adapt to reflect advances in the state of the art in capabilities and safety.

What is the standard of care for the reasonable development and deployment of AI agents today? Risk assessment, system documentation, appropriate use guidance, and deployment monitoring are industry norms for developers of all AI systems and are in many cases required by legislation including the EU AI Act. For AI agents in particular, design decisions and system affordances further human oversight. In part drawing from the prior literature (Shavit et al., 2023; Chan et al., 2024), the standard of care in developing AI agents to support human oversight could include:

- Risk assessment, system documentation,[13] appropriate use guidance, and deployment monitoring.
- Plan review and approval: prior to taking actions, AI agents surface plans to human users for review, modification, and approval.[14]
- Boundedness of deployment environment, accounting for certainty of that environment over time,[15] and ad-

[10] Chan et al. (2023) emphasizes four characteristics of agentic systems: underspecification of goals specified by users, directness of impact, goal-directedness, and long-term planning. The authors distinguish agenticness from autonomy, writing "While it is often an intuitive or useful description of a system, we find it combines distinct phenomena we wish to distinguish with our characteristics"(654), and go on to provide an example of an autonomous factory robot that performs bounded tasks. Shavit et al. (2023) echo the above, with a focus on goal complexity, environmental complexity, adaptability, and independent execution. Chan et al. (2024) simplifies the definition to focus on AI systems that "act directly in the world to achieve long-horizon goals" (5).

[11] Contrast this with much-criticized approaches to imposing liability and moral responsibility on AI systems directly. See Kolt (2024) footnote 207.

[12] Note, however, that absent adequate design affordances for oversight, users risk becoming "liability sponges" (Elish & Hwang, 2015) or "moral crumple zones" (Elish, 2019) taking the fall for AI failures. These risks are explored below in policy analysis on tort liability and consumer protection.

[13] Particularly detailing the models used in the system and a depiction of their orchestration architecture.

[14] This is current practice for at least some commercial systems, including GitHub Copilot Workspace. Looking to future advanced agents, Schulman (2023) has raised the idea of a "leash."

[15] Bounded deployments of agents within a particular application support oversight similar to use of any digital tool. Deploying agents capable of interacting on the internet at large or taking

herence to robots.txt.[16]

- Respect for humans and resources subject to AI interaction, including identification as an AI system.
- Approved list of tools and actions.
- Activity logging for traceability of actions.

The standard of care for reasonable deployment of AI agents would see users make use of these affordances. Efforts to collate practices and encourage greater transparency from developers across the development and deployment lifecycle are needed to improve understanding of the current standard of care.[17] Further research on human-AI interaction and safety (Vasconcelos et al., 2023; Park et al., 2024) for agents will similarly inform the standard of care. Policymaker statements, policy guidance, and enforcement can further understanding of appropriate oversight directly, by identifying and penalizing failures,[18] and indirectly, by incentivizing research that strengthens the standard of care over time. Policymakers have statutory and legal authority to begin this work today.

3. Existing policy on AI agents

3.1. The EU AI Act

Existing AI-specific laws, particularly the EU AI Act (Regulation (EU) 2024/1689), offer regulators levers for human oversight, including risk assessments, constrained deployment contexts, and monitoring of AI agents.[19] The AI Act imposes relevant obligations on general purpose AI models, high-risk AI systems, and certain systems warranting transparency.[20] The developers of general purpose AI models, including those that are integrated into AI agents, must provide documentation including intended and acceptable uses and evaluation criteria (Art. 53 and Annex XI). General purpose AI models that demonstrate capabilities across a wide range of tasks without specific training and while operating at a high level of autonomy can be classified as posing systemic risk (Annex XIII(e)); such models face risk assessment and mitigation requirements (Art. 55(b)).

High-risk AI system requirements additionally govern AI agents. Although general-purpose AI agents may not be high-risk AI systems per se, they could be deployed in regulated use cases intentionally, or unintentionally in cases of inadequate human oversight. These risks will likely see developers of AI agents and other general purpose AI systems technically guardrail and contractually limit the deployment contexts of AI systems to avoid high-risk-related obligations. Furthermore, users may demand such protections, because the EU AI Act regulates any third party as a developer if they modify the intended purpose of an AI system to be high risk.[21] In cases where systems are deemed high risk, they bring a number of requirements for risk assessment, monitoring, documentation, as well as human oversight "commensurate with the risks, level of autonomy and context of use" that includes both technical measures created by the developer and others to be implemented by the deployer before putting the system into service (Art. 14). Thus, AI regulation incentivizes constraints on AI agent actions and deployment environments that facilitate human oversight, and if such constraints are not used, AI agents may fall into high-risk AI categories that impose requirements directly.

Chapter IV of the AI Act introduces transparency requirements on AI systems interacting with human subjects and those generating synthetic content. AI systems interacting with a person must be identified as an AI system; similar state laws exist in the United States as well.[22] EU AI Office guidance could encourage appropriate disclosures that support monitoring, e.g., that could uniquely identify the system in question (Chan et al., 2024). More broadly, guidance from the EU AI Office and others can shape expectations for AI agent oversight and development. If this poses insufficient over time, the AI Act can be updated with new high-risk categories (Art. 112). As has been seen in the past, compliance with requirements could go well beyond jurisdictions where they may be enforced (Bradford, 2020).

3.2. Tort law

Regardless of whether an AI-specific law is applicable to a given system in a particular jurisdiction, legal expectations on responsible development and use exist today. In February 2024, a Canadian administrative tribunal held a company liable for the representations of its consumer-support chatbot, which had provided inaccurate advice to a grieving

actions in the physical environment could complicate oversight; bounding such deployments expressly by area or domain, number of actions, or otherwise could support oversight.

[16]The robot exclusion standard specifically with regards to agent exploration online at model inference time, not for training. See, e.g., the ChatGPT-User agent.

[17]E.g., Bavor & Taylor (2024).

[18]Note that reasonable oversight and agent affordances needed to support it may well depend on the specific facts or use case; see Crootof, et al. (2023) for related ideas on AI systems generally.

[19]While this analysis focuses on the EU AI Act, future work should consider Chinese regulations and US state laws. The Colorado Consumer Protections for AI law (2024) in particular shares similarities with the EU AI Act.

[20]Note that obligations on general purpose AI models with systemic risk, high-risk AI systems, and certain AI systems warranting transparency apply too to those available open source.

[21]Art. 25(1)(c). Note that "provider" is the EU AI Act term for developer.

[22]EU AI Act Art. 50(1) and Colorado Consumer Protections for AI law (2024) 6-1-1704 both apply to businesses; Cal. Bus. & Prof. Code §17941 applies to all persons in limited circumstances. Note that open source AI systems are subject to these provisions.

customer.[23] This reflects the broader legal status quo, which may chill adoption of AI agents: the user is liable for the actions taken by their AI system. A failure to adopt practices consistent with a standard of care and that result in harm could subject users and developers to claims of negligence with liability for damages. As noted in Section 2, greater transparency from developers can support stakeholders to inform and evolve this standard of care to mitigate foreseeable risks. Stakeholders can establish specific risks as foreseeable through public discussion, disclosures, and guidance. A full exploration of tort law is beyond this paper and its author. Despite a growing body of literature on liability and AI,[24] software liability in the United States has proved largely elusive for cases without physical harm (Choi, 2019). However, perhaps courts may be more likely to find fault in novel cases with AI agents.[25]

The EU recently updated product liability rules to apply to digital products. The Product Liability Directive (2024) gives consumers wide latitude to seek compensation for a range of harms, including psychological injury and corruption of data, caused by defective products.[26] The Directive, which must be transposed into member-state law in the next two years, states that non-compliance with safety requirements in other EU Law, including the AI Act, can lead to a presumption that the product is defective (Art. 7(1)).[27] A failure to update the AI agent with changes "necessary to maintain safety" similarly could confer liability (Art. 11(2)(c)). As with AI-specific laws, EU action on product liability may too influence AI development and deployment beyond the single market.[28]

[23] See, e.g., CBC coverage.

[24] Tort law for AI is an active area of research, though notably two leading scholars have published legal research centering human developers and users in liability for "risky [AI] agents without intentions" (Ayres & Balkin, 2024); for an incomplete sample of voluminous literature on the topic, see Kolt (2024) footnote 204.

[25] Indeed, in the Canadian case, the tribunal balked at the "remarkable" defendant claim that the chatbot was "a separate legal entity that is responsible for its own actions."

[26] Note that the companion AI Liability Directive was shelved in favor of the integrated Directive. Despite political agreement, the Directive has not yet been published in the Official Journal of the EU.

[27] This is further supported by Art. 10(4)(a) that establishes a presumption of causal link between product defectiveness and covered harm if "the claimant faces excessive difficulties, in particular due to technical or scientific complexity, in proving the defectiveness of the product or the causal link between its defectiveness and the damage, or both."

[28] Consider too, albeit more speculatively, the U.S. context for product liability. Unbounded or otherwise-designed AI agents that pose challenges for adequate human oversight could be considered defective: "a product is defective when, at the time of sale or distribution, it contains a manufacturing defect, is defective in design, or is defective because of inadequate instructions or warnings." Ayres & Balkan (2024) quoting Restatement (Third) of Torts, Product Liability (1997 ed.) §2(b). Design defects are proven by demonstrating the existence of a reasonable alternative design, presumably which, for a given AI agent, could include deterministic software and other AI systems with affordances for adequate oversight. Insofar as AI agents could pose risks of harm that rise to torts, product liability could incentivize developers to design AI systems such that they have adequate human oversight and to document them to that end. Further research should explore these possibilities.

3.3. Consumer protection

User liability for actions taken by AI agents may slow adoption of the technology. Companies will likely seek to overcome this hurdle by attempting to increase consumer trust, including by making safety claims and offering use guidance for appropriate human oversight.[29] To further consumer protection, regulators should set expectations for adequate human oversight and carefully scrutinize claims. In the U.S. the Federal Trade Commission (FTC) can take enforcement action to protect consumers. The FTC has already released guidance to companies "to keep your AI claims in check" underlining the need for accurate depiction of system capabilities (Atleson, 2024a) and for risk assessments and mitigations to consider foreseeable downstream uses as well as actual deployed use(Atleson, 2024b). Although the end of Chevron deference may complicate federal regulation of AI under existing law (Bullock, 2024), the FTC has clear remit to take action against unfair or deceptive trade practices, which would govern adoption-encouraging claims (15 U.S.C. §45(5)). Consumer protection authorities in additional jurisdictions could consider policy guidance for appropriate human oversight and documentation of AI systems including agents.

3.4. Cybercrime

The future widespread use of AI agents online could strain internet infrastructure and exacerbate cybersecurity risks. Web hosts may consider prohibiting AI agents in their terms of service, robots.txt, or otherwise technically restraining them via captcha[30] or login walls. Computer crime law gives these private restrictions criminal consequences, while also addressing hackers developing or script kiddies using AI agents pre-configured for malicious ends. The U.S. Computer Fraud and Abuse Act (CFAA; 18 U.S.C. §1030) outlaws accessing a computer resource "without authorization" intentionally and causing damage or defrauding. The CFAA has been used to prosecure developers and deployers of botnets.[31] The Department of Justice has previously issued

[29] They may also take on liability contractually or similarly providing indemnities as companies have done to spur adoption of generative AI tools amid copyright uncertainties.

[30] Acknowledging that captchas will need to continue to improve with increased AI capabilities. See, e.g., Arkose Bot Manager as a Time Magazine Best Invention of 2023.

[31] Berris (2023); e.g., see this 2024 Department of Justice indictment related to the 911 S5 Botnet.

policy guidance for bringing cases under the CFAA,[32] and could consider doing so again if AI agents pose harms in practice. Such guidance could underline that use of agents on the internet without adequate oversight (and that causes damage or strains resources) would violate the law.[33] This could have a chilling effect on both development and use of agents, much as the CFAA has chilled security research to-date.[34] More provocatively, prosecutors could ask if the development of an AI agent that cannot be adequately controlled to prevent harms at deployment, is this not essentially creating and trafficking in malware?

4. Discussion

Existing laws, user expectations, and enforcement actions constitute a political economy of deployment that shapes AI agent development and use. Multiple stakeholders can influence this political economy to chill unsafe autonomy. Policymakers, regulators, prosecutors, researchers, developers, and others can articulate and strengthen a standard of care for the reasonable development and deployment of AI agents subject to human oversight.

The standard of care can be advanced by an enforcement and research agenda. This effort can support systemic learning and societal adaptation as new autonomous capabilities are deployed and scrutinized gradually. It also incentivizes investment in AI safety and human-AI interaction research. This enforcement and research agenda would ultimately serve to retain human agency and accountability while respecting the existing global policy consensus for human oversight of AI systems.

The previous section began to articulate the enforcement agenda. The related research agenda to chill unsafe autonomy is expansive. Several directions are raised below:

- A survey of existing practices and interviews with leading developers could better clarify the current standard of care for AI agent development and deployment.
- Work could further specify the standard of care for specific facts and contexts, including deployment of agents capable of interacting on the internet and adherence to robots.txt.
- What could agent governance learn from scholarship on software security and efforts to articulate standards of care in that field?[35]
- Assessment of AI-specific laws could be expanded to cover regulatory proposals and to additional jurisdictions, including U.S. state law and Chinese regulations.
- How well do technical guardrails and contractual requirements work in practice to restrict high-risk use cases of AI systems today?
- How can regulatory sandboxes or other experimental policy tools be used to test and refine oversight mechanisms for AI agents?
- What may legal cases for defectiveness or negligence look like for the development and deployment of AI agents without adequate human oversight?
- Consumer protection authorities have existing authorities to penalize deceptive or unsafe AI systems. To what extent will the end of Chevron deference undermine the FTC's authority to prosecute these and deceptive trade practices, specifically?
- What should be learned from the history of FTC enforcement in emerging technology markets, including the adoption of privacy policies and related violations?
- What may draft policy guidance for CFAA enforcement related to AI agents look like, accounting for constraints imposed by the *Van Buren* (2021) ruling?
- Aside from the CFAA, what other cybercrime statutes in jurisdictions around the world may shape AI agent development and deployment?

The approach forwarded in this paper is not without limitations. First and foremost, Regulator attention is finite. Advocating for regulatory resources to be spent on AI agents before harms manifest may be politically infeasible and undermine enforcement against harms from other AI deployments. Other stakeholders may usefully produce research that supports agent governance enforcement, ideally aligned with broader government AI enforcement priorities, to reduce resources required for such actions when they are warranted. Second, regulatory scrutiny may not provide a strong-enough signal to steer agent development and deployment to chill unsafe autonomy; in other words, the standard of care may yield human oversight that while supposedly reasonable is wholly inadequate. Policy proposals could remedy this over time and should be tailored to reflect gaps in existing enforcement efforts. Third, the approach to

[32] See, e.g., this 2022 Department of Justice press release.

[33] Guidance could consider the use of an AI agent to interact with websites that require logins to violate prohibitions against intentionally accessing a protected computer and recklessly causing damage (§1030(5)(B)). The Supreme Court narrowed the CFAA in *Van Buren v. United States* (2021), complicating further use of this statute in cases where the individual can access a website and take actions not authorized by the provider. However, following this decision, many websites have restricted access behind login walls and now provide fact patterns that could provide for effective use.

[34] Though in this case for the worse: see Electronic Frontier Foundation (2020).

[35] See, e.g., Lawfare's project on Security by Design (Wittes & Rosenzweig, 2023).

chilling autonomy could bring risks of missed use, where autonomous AI agents could bring great benefits but are not deployed as widely because of regulatory scrutiny.

This workshop paper faces additional limitations. It has eschewed assessment of specific AI agents today. Detailed engagement with examples could benefit subsequent works. More broadly, market monitoring efforts for AI agent development and deployment could provide useful empirics to inform enforcement and governance research. The paper has focused on EU and, to some extent, US law. Further work can expand jurisdictions considered and depth of analysis. It has raised the standard of care concept but not explored how such an approach has been used previously and lessons from historical analogies to help illuminate foreseeable failure modes. These limitations can be addressed in further work, including by addressing questions raised above.

This paper has centered autonomy and its constraint, human oversight, as a foundation for AI agent governance. It analyzed existing policy tools that can be used to enforce a standard of care to adaptively constrain unsafe autonomy while steering development towards AI systems that preserve human agency and accountability.

Acknowledgements

Thanks to Noam Kolt, Merlin Stein, Mike Linksvayer, Alan Chan, Matthijs Maas, Shrey Jain, Aviya Skowron, and two anonymous reviewers for feedback on earlier drafts. Any errors are the author's alone.

References

OECD Recommendation of the Council on Artificial Intelligence, 2019. URL https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

The Bletchley Declaration by countries attending the AI safety summit, 2023. URL https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

Executive Order 14110 on the safe, secure, and trustworthy development and use of artificial intelligence, 2023. URL https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Hiroshima Process International Code of Conduct for Advanced AI Systems, 2023. URL https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems.

European Parliament legislative resolution of 12 March 2024 on the proposal for a directive of the European Parliament and of the Council on liability for defective products, 2024. URL https://www.europarl.europa.eu/doceo/document/TA-9-2024-0132_EN.html.

EU Artificial Intelligence Act, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689.

Colorado Consumer Protections for Artificial Intelligence, 2024. URL https://legiscan.com/CO/text/SB205/2024.

Seoul Ministerial Statement for advancing AI safety, innovation and inclusivity, 2024. URL https://www.gov.uk/government/publications/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024.

Atleson, M. Keep your AI claims in check, 2024a. URL https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check.

Atleson, M. The luring test: AI and the engineering of consumer trust, 2024b. URL https://www.ftc.gov/consumer-alerts/2023/05/luring-test-ai-and-engineering-consumer-trust.

Ayres, I. and Balkin, J. M. The law of AI is the law of risky agents without intentions. *University of Chicago Law Review Online*, 2024. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4862025.

Bavor, C. and Taylor, B. The agent development life cycle. *Sierra*, 2024. URL https://sierra.ai/blog/agent-development-life-cycle.

Berris, P. Cybercrime and the law: Primer on the computer fraud and abuse act and related statutes. CRS Report R47557, Congressional Research Service, Washington, DC, 2023. URL https://crsreports.congress.gov/product/pdf/R/R47557.

Bradford, A. *The Brussels effect: How the European Union rules the world*. Oxford University Press, 2020.

Bullock, C. What might the end of Chevron deference mean for AI governance? *Institute for Law & AI*, 2024. URL https://law-ai.org/chevron-deference.

CADMV. California autonomous vehicle regulations, 2024. URL https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/california-autonomous-vehicle-regulations/.

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 651–666, 2023.

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., et al. Visibility into AI agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 958–973, 2024.

Choi, B. H. Crashworthy code. *Wash. L. Rev.*, 94:39, 2019.

Cohen, M. K., Kolt, N., Bengio, Y., Hadfield, G. K., and Russell, S. Regulating advanced artificial agents. *Science*, 384(6691):36–38, 2024.

Crootof, R., Kaminski, M. E., Price, W., and Nicholson, I. Humans in the loop. *Vand. L. Rev.*, 76:429, 2023. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4066781.

Eastman, B., Collins, S., Jones, R., Martin, J., Blumenthal, M. S., and Stanley, K. D. A comparative look at various countries' legal regimes governing automated vehicles. *JL & Mobility*, pp. 1, 2023. URL https://repository.law.umich.edu/cgi/viewcontent.cgi?article=1019&context=jlm.

Electronic Frontier Foundation. Brief of amicus curiae computer security researchers, Electronic Frontier Foundation, Center for Democracy & Technology, Bugcrow, Rapid7, Sythe, and Tenable] in support of petitioner. Filed with Supreme Court of the United States, July 2020. URL https://www.eff.org/files/2020/07/08/19-783_eff_security_researchers_amici_brief_.pdf. No. 19-783, Van Buren v. United States.

Elish, M. and Hwang, T. Praise the machine! Punish the human! *Comparative Studies in International Systems, Working Paper*, (1), 2015.

Elish, M. C. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5, 2019.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., et al. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244*, 2024.

Horowitz, M. C. and Scharre, P. Meaningful human control in weapon systems. *CNAS*, 2015. URL https://www.files.ethz.ch/isn/189786/Ethical_Autonomy_Working_Paper_031315.pdf.

Huang, Q., Vora, J., Liang, P., and Leskovec, J. MLAgentBench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2023.

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. AI agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.

Kolt, N. Governing AI agents, 2024. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4772956.

Lessig, L. *Code: And Other Laws of Cyberspace*. Basic Books, 1999.

METR. Autonomy evaluation resources, 2024. URL https://metr.org/blog/2024-03-13-autonomy-evaluation-resources/.

Morris, M. R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. Levels of AGI: Operationalizing progress on the path to AGI. *arXiv preprint arXiv:2311.02462*, 2023.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., and Hendrycks, D. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

Robbins, S. The many meanings of meaningful human control. *AI and Ethics*, pp. 1–12, 2023.

Schulman, J. Keeping humans in the loop. Alignment Workshop, 2023. URL https://www.alignment-workshop.com/nola-talks/john-schulman-keeping-humans-in-the-loop.

Securities and Exchange Commission. Staff report on algorithmic trading in U.S. capital markets. Staff report, U.S. Securities and Exchange Commission, August 2020. URL https://www.sec.gov/files/algo_trading_report_2020.pdf.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., et al. Practices for governing agentic AI systems. 2023. URL https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

Simmler, M. and Frischknecht, R. A taxonomy of human–machine collaboration: Capturing automation and technical autonomy. *AI & Society*, 36(1):239–250, 2021.

Simmons-Edler, R., Badman, R., Longpre, S., and Rajan, K. AI-powered autonomous weapons risk geopolitical instability and threaten AI research. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

UK AI Safety Institute. Agents, 2024. URL https://ukgovernmentbeis.github.io/inspect_ai/agents.html.

U.S. Copyright Office Review Board. Re: Second request for reconsideration for refusal to register a recent entrance to paradise, 2022. URL https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., and Krishna, R. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):1–26, 2024.

Wareham, M. Stopping killer robots, 2020. URL https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and.

Wex. Standard of care, 2021. URL https://www.law.cornell.edu/wex/standard_of_care.

Wittes, B. and Rosenzweig, P. Announcing a new lawfare project on 'security by design', 2023. URL https://www.lawfaremedia.org/article/announcing-a-new-lawfare-project-on-security-by-design.

Yang, G.-Z., Cambias, J., Cleary, K., Daimler, E., Drake, J., Dupont, P. E., Hata, N., Kazanzides, P., Martel, S., Patel, R. V., et al. Medical robotics–regulatory, ethical, and legal considerations for increasing levels of autonomy. *Science Robotics*, 2(4):eaam8638, 2017.

Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., and Press, O. Swe-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint arXiv:2405.15793*, 2024.

Zittrain, J. We need to control AI agents now. *The Atlantic*, July 2024. URL https://www.theatlantic.com/technology/archive/2024/07/ai-agents-safety-risks/678864/.