Chilling autonomy: Policy enforcement for human oversight of AI agents

1. Introduction

AI research and development is increasingly focused on AI agents. In contrast to dominant applications of large language 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 their environment over time (Wang et al., 2024). Adoption of AI agents is a function not only of their underlying performance but also of legal consequences. This paper takes first steps to define the role that policy stakeholders have to play today in offering interpretive guidance for existing law and otherwise applying scrutiny to developers of and the market for AI agents, in order to ensure that systems do not pose risks that would be democratically unacceptable. The analysis focuses on an established expectation for AI systems: that they be developed for and deployed with adequate human oversight.

Today AI agents are largely marketing aspiration, though tool-use and scaffolding software on large language models have enabled systems of modest capabilities in research settings (Yang et al., 2024; Huang et al., 2023) that often neglect costs (Kapoor & Narayanan, 2024). Capital and research investments will likely yield AI agents that can plan over longer time horizons and take an increasing range of complex actions in increasingly complex environments, including interacting with the internet at large. As planning, actions, and deployment environments all increase in complexity, the stakes rise for ensuring adequate human oversight. Risks from not-yet-developed capable AI agents are already influencing policy. The Biden Executive Order on AI (2023), Hiroshima Process International Code of Conduct (2023), and other documents emphasize risks from loss of "human control or oversight," systems that "substantially" 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 automated completion across multiple malicious tasks. AI Safety Institute and other evaluations evaluate performance across chains of specific tasks (METR, 2024). More broadly, the successful completion of chains of tasks by AI agents present underexplored implications for labor automation and the future of work (Eloundou et al., 2023).

2. Centering human oversight

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.[1] This literature has surfaced worthwhile questions for developers to consider in designing AI systems (Shavit et al., 2023) and presented some design concepts that could support governance (Chan et al., 2024). Another study looks to the economic theory and law of agency relationships for governance inspiration, taking the perspective that AI agents mark a paradigm shift for AI systems as 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, AI development is happening within a policy context today, which does not recognize AI systems as anything other than tools built and used by responsible humans. Human oversight and the level of autonomy (Simmler & Frischknecht, 2021; Yang et al., 2017) permitted in use, not agenticness, are more useful concepts to understand how such systems are and will be governed. As capabilities improve, systems may effectively function with greater autonomy, and this could change inherent needs for and forms of human oversight.[2] Yet, any such changes would need to be negotiated within existing law and the political economy of deployment.[3]

[1] 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).

[2] 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)

[3] For example, consider expectations from users in adopting agents and providers of websites and other services subject to agents-environment interaction. Norms and market demands, alongside law and capabilities will shape agent development and deployment (Lessig, 1999).

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). The EU AI Act (2024) dedicated an article to it (analyzed below). Laws globally expect humans to be in the loop for intellectual property protections to apply.[4] Efforts continue to ban lethal autonomous weapons (Welsh, 2019). Autonomous vehicles are deployed with express permission from regulators (CADMV, 2024).

The development of autonomous AI agents challenges this policy consensus in principle. Policymakers, regulators, prosecutors, and other stakeholders should take steps to clarify expectations for adequate human oversight in AI agent development and deployment. Absent this clarity, users risk becoming "liability sponges" (Elish & Hwang, 2015) or "moral crumple zones" (Elish, 2019) taking the fall for AI failures because the adequacy of oversight is context dependent (Crootof et al., 2023). Further research on human-AI interaction and safety (Vasconcelos et al., 2023; Park et al., 2024) for agents is needed. In the meantime, policymaker vigilance and enforcement can further understanding of appropriate oversight directly, by identifying and penalizing failures, and indirectly, by incentivizing research that advances a duty of care. Policymakers have statutory and legal authority to begin this work today.

3. Policy analysis

Existing AI-specific laws offer regulators levers for appropriate human oversight, constrained deployment contexts, and monitoring of AI agents. The EU AI Act (2024) and the Colorado Consumer Protection for AI law (2024) scrutinize AI systems that could be used in consequential decision-making or other high-risk settings. Although general-purpose AI agents may not be per se high-risk AI systems, they could be deployed in such cases intentionally, or unintentionally in cases of inadequate human oversight.[5] Such risks will likely see AI developers both technically guardrail and contractually limit the deployment contexts of AI agents to avoid high-risk-related obligations. Furthermore, consumers 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.[6] 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 (Article 14). Thus, AI regulation incentivizes constraints on AI agent activity that facilitate human oversight, and if such constraints are not used, AI agents may fall into high-risk AI categories that impose requirements that further human oversight.

Guidance from the EU AI Office and others can shape expectations for AI agent oversight and development. If this poses insufficient over time, the EU AI Act can be updated with new high-risk categories (Article 112). More generally, AI systems interacting with a person must be identified as AI systems under the EU AI Act, Colorado Consumer Protection for AI law, and California state law.[7] Guidance could similarly encourage appropriate disclosures that support monitoring, e.g., that could uniquely identify the system in question (Chan et al., 2024).

4. Contribution

This paper is the beginning of a larger enforcement and research agenda. It centers AI agent governance in existing policy, grounding discussion in autonomy and human oversight. The full paper will consider additional policy areas: liability, consumer protection, and cyber-crime law. These, together with existing AI-specific laws, provide regulators, policymakers, prosecutors, and other stakeholders tools to steer AI development towards agents that enable appropriate human oversight in line with shared global values and commitments. This enforcement effort will incentivize good practices and advance needed investment in AI safety and human-AI interaction research.

References

OECD recommendation of the council on artificial intelligence, 2019. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

The Bletchley declaration by countries attending the AI safety summit, 2023. Available: https://www.gov.uk/government/publications/ai-

[4] E.g., the U.S. Copyright Office (2022) rejected protection for an autonomously created artwork and the UK did similarly in a patent case (2023).

[5] The EU AI Act expressly accounts for this (discussed below) and the Colorado Consumer Protections for AI law excludes chatbots but not general-purpose AI systems, particularly those capable of taking actions in environment, from possible high-risk scope 6-1-1701(9). Note that open source AI systems are in scope of high-risk provisions.

[6] Article 25(c). Note that "provider" is the EU AI Act term for developer.

[7] EU AI Act Article 50(1) and CO 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.

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. Available: 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. Available: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems.

EU artificial intelligence act, 2024. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01$_EN.pdf$.

Colorado consumer protections for artificial intelligence, 2024. Available: https://legiscan.com/CO/text/SB205/2024.

Seoul ministerial statement for advancing AI safety, innovation and inclusivity, 2024. Available: 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.

BBC. AI cannot patent inventions, uk supreme court confirms, 2023. Available: https://www.bbc.com/news/technology-67772177.

Board, U. C. O. R. Re: Second request for reconsideration for refusal to register a recent entrance to paradise, 2022. Available: https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf.

CADMV. California autonomous vehicle regulations, 2024. Available: 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.

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. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract$_id$ = 4066781.

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.

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. and Narayanan, A. AI leaderboards are no longer useful. it's time to switch to pareto curves, 2024. AI Snake Oil, available: https://www.aisnakeoil.com/p/ai-leaderboards-are-no-longer-useful.

Kolt, N. Governing AI agents, 2024. Available at SSRN: 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. Available: 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.

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. *Research Paper, OpenAI*, 2023.

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

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.

Welsh, S. Regulating lethal and harmful autonomy: Drafting a protocol vi of the convention on certain conventional weapons. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 177–180, 2019.

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.