Position: AI Agents & Liability – Mapping Insights from ML and HCI Research to Policy

Anonymous Author(s)
Affiliation
Address
email

Abstract

AI agents are loosely defined as systems capable of executing complex, open-ended tasks. Many have raised concerns that these systems will present significant challenges to regulatory/legal frameworks, particularly in tort liability. However, as there is no universally accepted definition of an AI agent, concrete analyses of these challenges are limited, especially as AI systems continue to grow in capabilities. In this paper, we argue that by focusing on *properties* of AI agents rather than the threshold at which an AI system becomes an agent, we can map existing technical research to explicit categories of "foreseeable harms" in tort liability, as well as point to "reasonable actions" that developers can take to mitigate harms.

1 Introduction

AI agents are loosely defined in literature as AI systems capable of independently pursuing complex goals. Existing systems, like AutoGPT [48, 4], are being enhanced with more autonomy, and future agents will likely plan farther ahead, adapt and act in more complex settings. As AI agents promise to support humans across a wide range of tasks, the associated reduction in human control/oversight introduces notable risks and uncertainties for our legal system [18, 29]. While there is increasing interest in understanding how legal and regulatory frameworks apply to the governance of AI agents [28, 34, 12, 9, 41], the lack of a universally accepted definition of 'AI agent' limits concrete analysis.

In this work, we argue that by focusing on *key properties* of current and future AI systems, rather than the threshold at which an AI system becomes an agent, we can already leverage Machine Learning (ML) and Human-Computer Interaction (HCI) literature to provide insights on key legal questions. Here, we focus on two questions in fault-based liability[1], which imposes a general duty of care to avoid intentionally or negligently causing harm to others: "who is best able to prevent harms?" and "what is a reasonable duty of care?"

In our analysis, we study on three properties (autonomy, complexity and adaptability) that measure the capabilities AI agents; we organise relevant actors for liability using the AI value-chain (consisting of foundation model developers upstream, application developers midstream and end-users downstream). We argue that the multitude of actors in the AI value-chain, along with the increasing capabilities of agents, makes it difficult to determine which actors control the outcomes of agent actions. Addressing these challenges, we draw on ML and HCI research to identify harms that can arise due to the

[1] Some jurisdictions have passed or proposed specific legislation on liability for AI harms. For example, the recently updated EU Product Liability Directive (European Parliament, 2024) now explicitly includes software/AI systems; EU's proposed AI Liability Directive and California's Bill SB1047 all target AI systems. However, traditional theories of liability still apply in most contexts.

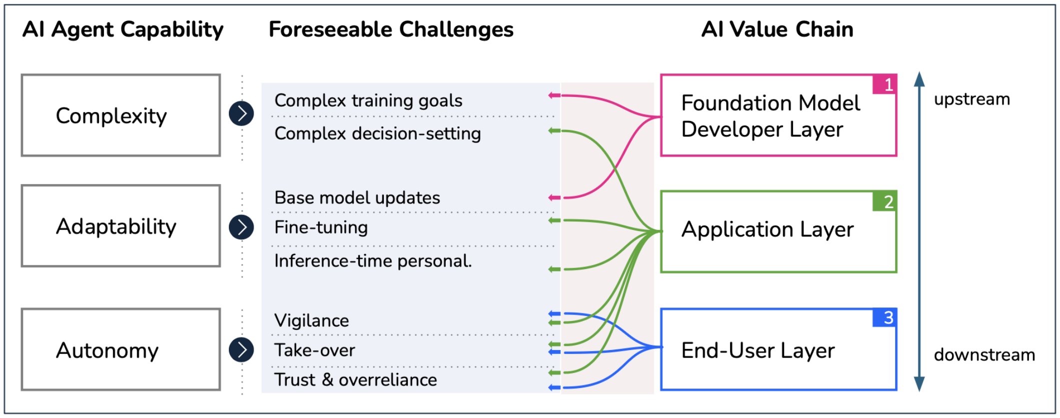

Figure 1: We map actors in the AI value chain to foreseeable challenges. Increasing agent capabilities leads to challenges (e.g. autonomy can lead to users overrelying on AI), and different actors in the chain have different degrees of control over challenges (e.g. HCI research implies that over-reliance is foreseeable by application developers). This can help define duties of care for different actors.

autonomy, complexity and adaptability of agents; we then identify actors who can mitigate these harms (Figure 1). Together, these insights can help define duties of care along the AI value-chain [2].

1.1 Key Definitions

What is the AI Value Chain? We define the AI Value Chain as the sequence of steps leading to the deployment of an AI system in a specific setting. We organise the steps, as well as the actors, into layers that describe the changes made to the system. Consider the example of an AI agent deployed in a hospital that summarises patient health records for clinicians. At the Foundational Model Developer Layer, actors are developers who train and define the general capabilities of foundation models. In our example, the agent may be built upon Meta's Meditron, a suite of open-source medical Large Language Models [10]. At the Application Layer, actors are developers who build infrastructures that allow AI systems to interface with the end-user and the environment. In our example, a third-party company may adapt the base-model, they may also build a system for the modified model to interface with hospital workers. At the End-User Layer, actors delegate tasks to an agent or use information from an agent to accomplish a specific goal. In our example, end-users are clinicians who make decisions for patient care informed by the summaries produced by the AI agent.

In this paper, we focus on analysing which actors in the AI Value Chain can foresee and mitigate harms that arise when the final system is deployed. In our example, consider when the AI agent produces an inaccurate summary that leads to a patient receiving an incorrect treatment and thus suffering a negative health outcome. Could the developers upstream and the clinicians downstream have foreseen this harm? What could each set of actors have done to prevent it?

What is liability law? Broadly, legal liability can arise from many areas of law, including contract law, criminal law, consumer protection law and tort law. We focus on tort liability[3], which, in part, aims to ensure that victims of harmful actions are compensated, and that those responsible for causing harm are held responsible. Thus, tort liability is about corrective justice (ensuring that victims are adequately remedied), and about deterrence/harm prevention (motivating people to be careful because they can be held liable for harms they cause [46]). Based on such theories, tort liability should be placed with the person who, if acting reasonably, is able to prevent the harm.[4]

[2]We acknowledge that different jurisdictions have distinct tort law traditions, statutes, and case law. This paper does not aim to discuss specific legal solutions for particular jurisdictions, but instead aims for a high-level discussion on liability that may be relevant in multiple jurisdictions. We also recognize the existing legal scholarship on tort law and AI - for a comprehensive overview see footnote 5 of [25]

[3]Mentions of 'liability' in this paper should thus be read as 'tort liability'.

[4]From a policy perspective there can also be other motivations for placing tort liability with a certain actor. For example, an actor with 'deep pockets' who can ensure compensation.

Both common law and civil law systems have 'theories of liability', developed through a combination of jurisprudence (decisions by judges) and statutes (laws). Some common theories of liability include fault liability (which includes intentional torts or negligent torts), strict liability (liability not requiring 'fault', usually for dangerous activities or goods), product liability (liability of manufacturers for defective products), and vicarious liability (liability for conduct of others). We focus on fault liability, in particular on negligence, although the lessons from this paper may also be relevant for other forms of tort liability or legal liability. Fault-based liability is the most general form of liability; it applies even in the absence of specific legislation. For the purposes of this paper, we are focusing on unintentional harms caused by AI agents, i.e. *negligent torts*.[5]

What is negligence? A 'tortfeasor' (person responsible for the harm) can be held liable for a negligent tort when they fail to take reasonable action to prevent a foreseeable harm. Negligence hinges on the concept of a breached 'duty of care', meaning that the potential tortfeasor did not take the reasonable care expected of them in a certain situation (e.g. failing to set up a warning sign when leaving open a hatch in the middle of a pedestrian walkway), and this failure leads to harm.

The test is not whether that specific tortfeasor had foreseen the harm (that would absolve oblivious but culpable tortfeasors), but whether a 'reasonable person' in the position of the tortfeasor would have foreseen that this harm could happen. This 'reasonableness standard' translates into a 'duty of care'. The duty of care is an objective standard informed by the actions of others: industry standards and best practices, academic research, statements by policymakers, and legal requirements [11].

What's Challenging in Applying Existing Tort Liability Rules to AI Systems? Value chains of AI systems tend to be complex, with many actors involved in different aspects of system development and deployment [5]. Under the current status quo, liability concentrates downstream towards the end-user [47, 11]. This may be problematic, as downstream actors are small players and may be less able to shoulder the liability compared to upstream big tech developers. Furthermore, downstream actors may have less expertise, capacity or ability (e.g. access to base-models) to meet the requisite duty of care [13]. Finally, AI systems may raise new risks, like immaterial harms that can result from social biases baked into AI systems, that are not currently addressed by tort liability [8, 25].

What is an Agent? AI agents have been defined in different ways in literature. Generally speaking, definitions center around the capability of agents: such systems would have the ability to perceive and operate in complex environments, and to autonomously adapt their strategies and actions based on new input [35, 21, 40, 23, 15]. In this paper, we define an AI agent as *an AI system that is deployed in a real-life decision making setting*, and we focus on the capability of these system as measured by autonomy, adaptability and complexity. We focus on these properties as they significantly challenge humans' ability to anticipate/prevent the harmful outcomes of AI agent actions.

2 Diffusion of Liability Along the AI Value Chain

We describe specific ways the autonomy, adaptability and complexity of AI agents may lead to harmful outcomes. Furthermore, for each actor in the AI value chain, we draw from current research to identify the degrees to which they may foresee and mitigate harms (summary in Figure 1).

Autonomy. We focus on two levels of autonomy when AI agents interact with human users: (1) AI agents that support human decision-making, and (2) AI agents that operate under human supervision.

AI agents with low autonomy include current decision-support tools that rely on humans for critical assessment of AI outputs. For example, AI agents have been deployed to support clinical decision making, such as by performing risk assessments of patients and recommending treatment [38]. Clinicians using these systems need to determine whether or not the AI recommendations are valid or useful. However, literature in Human-Computer Interaction (HCI) demonstrates that humans have a strong propensity to over-rely on and over-trust AI assistants [22, 6].

For AI agents with greater degrees of autonomy, a common design choice is to cast human users in supervisory roles, monitoring the system for errors and taking over during exceptional circumstances [3]. In a clinical setting, this may look like an agent that automatically screens mammography for cancer, supervised by a clinical expert who monitors the system and steps in to perform manual

[5] Liability is relatively straightforward for intentional harms, as it will usually be the person intentionally causing the harm who 'controlled the outcome' and is liable.

diagnosis in cases of errors and exceptions. Unfortunately, a large body of work in HCI shows us that humans struggle with vigilance (i.e. sustaining attention while performing monotonous tasks over long periods of time), thus making them poor supervisors of automated systems [1, 42, 37]. In exceptional cases where humans are required to take-over for the AI agent, the problem of vigilance amplifies the difficulty of control transfer: the human, having reduced prior involvement with the task, is likely to struggle with reacting appropriately and in a timely fashion to the situation [31].

What are 'foreseeable harms': Overwhelmingly, literature on human factors in computing teaches us that naïve integration of agents into AI+human teams can exacerbate human error [2, 50]. That is, existing HCI research can help establish explicit and concrete categories of 'foreseeable harms', as well as point to 'reasonable actions' application developers can take to mitigate these harms.

Who can prevent these harms: Design of human+AI interfaces happens in the Application Layer, where developers can leverage literature to address problematic ways end-users will interact with agents. For example, AI systems that provide continuous-support (rather than recommendation-centric support) have been shown to help users maintain situational awareness in human+AI teams [52].

Adaptability. We focus on two common ways that agents adapt: (1) changing the system's base-model at pre-deployment or model-update time, (2) personalising system output with specialised input at inference (i.e. decision-making) time.

Many foundation models can be customised. Outside of open-source models, major developers like OpenAI, Google, Microsoft, Meta, Anthropic, and Amazon have existing or proposed mechanisms for downstream developers to adapt their models for specific applications through fine-tuning [33]. That is, these models can be updated based on new data and interactions in order to learn new skills. However, unlike traditional software updates, which are designed to preserve existing functionality whilst adding new ones, updating AI models with new data can lead to (1) degradation of existing functionality [51] (for example, safety guardrails can be quickly by-passed with fine-tuning, providing user access to dangerous information); (2) unexpected new biases and failure modes, as new data interacts with existing ones on which the model was trained [45].

Even when models in AI agents are fixed, the outputs of these systems can be personalised to individual or groups of end-users at inference-time, by including special information in system inputs. For example, language models can suggest personalised email subjects when shown user's past emails [36]. However, personalising model outputs risks confirmation bias (selective reinforcement of users' existing opinions) [39, 14]; it can even result in behaviours that can be categorised as deceptive, e.g. actively steering users away from or hiding contradictory information [20]. Furthermore, and perhaps surprisingly, personalisation can harm model performance. That is, personalising a model to a specific group of users can lower the model's performance at a group level [43, 53]. For example, providing gender to language models when generating a recommendation letter can increase model hallucination as well as diminish language associated to "excellence" [44].

What are 'foreseeable harms': Research shows that, for adaptable agents, anticipating harms based on pre-deployment functionalities cannot cover the range of harms that may result from post-deployment changes. Existing works that study AI models in the regimes of fine-tuning, continual learning and transfer learning can already help us anticipate emergent agent behaviours and associated risks.

Who can prevent these harms: How much and in what ways an agent can adapt are design choices made both at the Foundation Model Developer Layer and the Application Layer. While safe and effective adaptation is an open research question, upstream developers have a responsibility to test agents for failures known to arise in adaptation, and disclose these risks to downstream actors.

Complexity. We focus on two aspects of complexity: (1) complexity of the agent's goals (specifically, its training objective), and (2) complexity of the agent's decision setting (specifically, the length of the planning horizon and the effective size of the environment).

Behaviours of AI agents are often determined by multiple objectives that may be in tension. For example, foundation models trained with human feedback implicitly balance potentially conflicting preferences of different users. Furthermore, there are tensions between social welfare goals (e.g. safety) and personalisation goals (e.g. open access to information for individual users). However, agents are often trained by maximising a single combination of multiple objectives, without explicitly managing trade-offs amongst them. Thus, the resulting agent can make unexpected and undesirable compromises [19], for example, by sacrificing safety in order to align with user preferences.

In addition to the multiple objectives, models in AI agents are frequently trained with "blackbox" and underspecified objectives. For example, foundation models are often fine-tuned with direct human feedback on model outputs [27], or with datasets of human annotated examples that encode human preference [24]. The objective implied by human feedback and annotated examples is not defined in closed-form, and hence cannot be directly inspected or interrogated. Furthermore, given a set of examples, there are often multiple plausible objectives that a model can infer that would cause it to generalize in dramatically different ways. Thus, training with implicit and underspecified objectives often leads to unexpected and undesirable agent behaviours [30, 26]. For example, an agent trained to interact with humans in more naturalistic and context-sensitive ways can learn, as unintended side-effects, to hold strong political opinions and to pursue potentially dangerous "subgoals" (e.g. to accumulate resources for the pursuit of current goal) [32].

Even when training objectives of AI agents can be validated, it is hard to design interfaces that allow end-users to anticipate outcomes of agent actions, when actions take place over long planning horizons and in open environments. Existing works show that humans can even struggle to understand decisions of simple AI systems in single-shot decision settings [7]. As the planning horizons of agents increase, and as the size of the environment as well as the number of other agents in the environment grow, the sequence of agent decisions becomes too complex to directly inspect. For example, an open question is how to effectively summarise complex policies for sequential decision-making: simply enumerating the agent's actions over the large number of possible states of the environment yields results that are uninterpretable to end-users [49, 16].

What are 'foreseeable harms': A number of failures of AI models can be exposed by quantifying how trained models trade-off different task-relevant desiderata. There is also a large body of literature that can surface biases models learn from human data [17]. Finally, research on explainable AI (XAI), as well as HCI, can anticipate failures of Human+AI teams in complex decision settings.

Who can prevent these harms: The complexity of an agent's goals can be determined at the Foundational Model Developer Layer (where the capabilities of the base-model is determined) as well as at the Application Layer (where capabilities may be added). Upstream developers have a responsibility to explicitly prevent models from making undesirable trade-offs. The complexity of an agent's decision setting is primarily determined at the Application Layer. Application developers need to design Human+AI interfaces guided by best practices in XAI and HCI.

3 Implications for Policy & Technical Research

Increasing capabilities of AI agents challenge human control when these systems are deployed in real-life. However, we argue that existing research in ML and HCI point to ways that actors in the AI value chain can already better foresee and mitigate potential harms. From the perspective of fault-based liability, actors at the *Foundation Model Developer Layer* have a responsibility to disclose models' training objectives and understand trade-offs models make between different objectives. They have a responsibility to check for and mitigate known model biases that arise from data selection and training; they should also test for how biases affect models in common down-stream tasks. Safe-guards should be implemented against inappropriate and potentially unsafe types of model customisation. At the *Application Layer*, when customising models, developers have a responsibility to test for (and address) known model failures due to fine-tuning, as well as for unequal model performance when it is personalised to different end-users. Furthermore, when integrating models into Human+AI systems, they have a responsibility to anticipate/address known challenges in human-AI interactions. At the *End-User Layer*, users should understand the types of tasks the AI agent can safely perform. They should especially monitor the agent's behaviour when it is used in a new setting.

Upstream actors have a general responsibility to expose conditions under which the model has been tested, and how it scored on different evaluations. Transparency around evaluation settings helps downstream actors perform due diligence when choosing to deploy AI agents for a specific use-case.

Call for interdisciplinary research. We see a need for more interdisciplinary research between core ML and HCI, to concretely connect properties of models (e.g. catastrophic forgetting, interference, reward hacking, pathologies arising from multi-objective optimization) to specific impacts on users' abilities to anticipate/mitigate harms of AI systems. We also see a need for more collaborations between policy/law makers and technical researchers to formalise regulatory/legal principles as technical desiderata (e.g. training procedures, objectives and metrics, socio-technical evaluations).

Impact Statement

We argue that interdisciplinary collaboration between technical (especially between ML and HCI) researchers and policy/law makers is necessary for effectively and responsibly addressing emerging societal challenges due to complex AI systems. By involving ML and HCI researchers in policy/legal processes, policy and lawmakers can better understand capabilities and limitations of emerging technologies, enabling the development of regulations that are informed and effective. For the technical community, engagement with law/policy professionals opens avenues for shaping regulatory environments in which they operate. Collaborating with legal/policy experts ensures that ML systems are designed with compliance in mind, reducing the risk of costly redesigns or legal challenges. It also encourages the adoption of ethical best practices, fostering public trust in AI applications.

A Appendix: Recommendations for Policymakers

As AI agents grow in autonomy, complexity, and adaptability, assigning and determining liability for harms becomes increasingly challenging. In this paper, we have shown that interdisciplinary research, particularly at the intersection of Machine Learning (ML), Human-Computer Interaction (HCI), and law, can help us establish clearer liability frameworks. Achieving this, we believe that there are concrete actions that policymakers can take to create the foundation for more robust and nuanced research into AI liability. We recommend the following:

Enhance visibility into AI agent development and deployment

- Facilitate research access to agent usage data to better understand human-agent interactions, similar to research access provisions under the Digital Services Act (DSA)
- Require identification systems for agents [**?**] and logging mechanisms to verify AI agents' actions
- Mandate stage- and context-specific evaluations and audits throughout the AI agent lifecycle through comprehensive auditing regimes
- Establish mandatory incident reporting and information sharing protocols

Strengthen technical governance capacity to support judicial assessment of liability

- Ensure sufficient socio-technical expertise within regulatory agencies through targeted hiring, training programs, and continuous education initiatives
- Explore flexible models such as scientific advisory boards or fellowship programs to source and integrate external expertise
- Develop standardized methodologies for assessing AI system capabilities and limitations in legal contexts

Establish and enforce a comprehensive duty of care for AI development and deployment

- Incentivize the development and adoption of industry-wide standards, particularly focusing on safety measures and trust infrastructures
- Support research to concretely connect properties of AI models (e.g., catastrophic forgetting, interference, reward hacking) to legal liability frameworks
- Facilitate the public dissemination of new safety procedures and formally promulgate these standards through regulatory channels

References

[1] Julia Adler-Milstein, Donald A Redelmeier, and Robert M Wachter. The limits of clinician vigilance as an ai safety bulwark. *Jama*, 2024.

[2] Saar Alon-Barkat and Madalina Busuioc. Human–ai interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1):153–169, 2023.

[3] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[4] Timo Birr, Christoph Pohl, Abdelrahman Younes, and Tamim Asfour. Autogpt+ p: Affordance-based task planning with large language models. *arXiv preprint arXiv:2402.10778*, 2024.

[5] Ian Brown. Allocating accountability in ai supply chains, 2023.

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

[7] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.

[8] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, 2023.

[9] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, et al. Visibility into ai agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 958–973, 2024.

[10] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[11] Peter Cihon. Chilling autonomy: Policy enforcement for human oversight of ai agents. In *41st International Conference on Machine Learning, Workshop on Generative AI and Law*, 2024.

[12] Michael K Cohen, Noam Kolt, Yoshua Bengio, Gillian K Hadfield, and Stuart Russell. Regulating advanced artificial agents. *Science*, 384(6691):36–38, 2024.

[13] Connor Dunlop. Regulating ai foundation models is crucial for innovation, 2023.

[14] Sharon Ferguson, Paula Akemi Aoyagui, Young-Ho Kim, and Anastasia Kuzminykh. Just like me: The role of opinions and personal experiences in the perception of explanations in subjective decision-making. *arXiv preprint arXiv:2404.12558*, 2024.

[15] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244*, 2024.

[16] Jasmina Gajcin, Rahul Nair, Tejaswini Pedapati, Radu Marinescu, Elizabeth Daly, and Ivana Dusparic. Contrastive explanations for comparing preferences of reinforcement learning agents. *arXiv preprint arXiv:2112.09462*, 2021.

[17] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024.

[18] Sara Gerke, Timo Minssen, and Glenn Cohen. Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare*, pages 295–336. Elsevier, 2020.

[19] Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

[20] Betty Li Hou, Kejian Shi, Jason Phang, James Aung, Steven Adler, and Rosie Campbell. Large language models as misleading assistants in conversation. *arXiv preprint arXiv:2407.11789*, 2024.

[21] Qiuyuan Huang, Naoki Wake, Bidipta Sarkar, Zane Durante, Ran Gong, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Noboru Kuno, Ade Famoti, et al. Position paper: Agent ai towards a holistic intelligence. *arXiv preprint arXiv:2403.00833*, 2024.

[22] Sarah Jabbour, David Fouhey, Stephanie Shepard, Thomas S Valley, Ella A Kazerooni, Nikola Banovic, Jenna Wiens, and Michael W Sjoding. Measuring the impact of ai in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *Jama*, 330(23):2275–2284, 2023.

[23] Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. Ai agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.

[24] Y Kim, X Xu, D McDuff, C Breazeal, and HW Park. Health-llm: large language models for health prediction via wearable sensor data, arxiv. *arXiv preprint arXiv:2401.06866*, 2024.

[25] Noam Kolt. Governing ai agents. *Available at SSRN*, 2024.

[26] Victoria Krakovna. Specification gaming examples in AI, 2018.

[27] Hengli Li, Song-Chun Zhu, and Zilong Zheng. Diplomat: a dialogue dataset for situated pragmatic reasoning. *Advances in Neural Information Processing Systems*, 36:46856–46884, 2023.

[28] Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, and Anca Dragan. Learning to model the world with language. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29992–30017. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/lin24g.html`.

[29] John D McGreevey, C William Hanson, and Ross Koppel. Clinical, legal, and ethical aspects of artificial intelligence–assisted conversational agents in health care. *Jama*, 324(6):552–553, 2020.

[30] Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. Feedback loops with language models drive in-context reward hacking. *arXiv preprint arXiv:2402.06627*, 2024.

[31] Samir Passi and Mihaela Vorvoreanu. Overreliance on ai literature review. *Microsoft Research*, 2022.

[32] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

[33] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

[34] Maria Abi Raad, Arun Ahuja, Catarina Barros, Frederic Besse, Andrew Bolt, Adrian Bolton, Bethanie Brownfield, Gavin Buttimore, Max Cant, Sarah Chakera, et al. Scaling instructable agents across many simulated worlds. *arXiv preprint arXiv:2404.10179*, 2024.

[35] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

[36] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023.

[37] Abigail Sellen and Eric Horvitz. The rise of the ai co-pilot: Lessons for design from aviation and beyond. *Communications of the ACM*, 67(7):18–23, 2024.

[38] Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR medical informatics*, 8(7):e15182, 2020.

[39] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

[40] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic ai systems. *Research Paper, OpenAI, December*, 2023.

[41] Julia Smakman. Ai assistants: Helpful or full of hype?, 2024.

[42] Helen Smith and Kit Fotheringham. Artificial intelligence in clinical decision-making: rethinking liability. *Medical Law International*, 20(2):131–154, 2020.

[43] Vinith Menon Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms performance: reconsidering the use of group attributes in prediction. In *International Conference on Machine Learning*, pages 33209–33228. PMLR, 2023.

[44] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

[45] Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. On negative interference in multilingual models: Findings and a meta-learning treatment. *arXiv preprint arXiv:2010.03017*, 2020.

[46] Ernest J Weinrib. Deterrence and corrective justice. *UCLA L. Rev.*, 50:621, 2002.

[47] Maria Yagoda. Airline held liable for its chatbot giving passenger bad advice—what this means for travelers. *URL https://www. bbc. com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know*, 2024.

[48] Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*, 2023.

[49] Jiayu Yao, Sonali Parbhoo, Weiwei Pan, and Finale Doshi-Velez. Policy optimization with sparse global contrastive explanations. *arXiv preprint arXiv:2207.06269*, 2022.

[50] Hubert D Zając, Dana Li, Xiang Dai, Jonathan F Carlsen, Finn Kensing, and Tariq O Andersen. Clinician-facing ai in the wild: Taking stock of the sociotechnical challenges and opportunities for hci. *ACM Transactions on Computer-Human Interaction*, 30(2):1–39, 2023.

[51] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

[52] Zelun Tony Zhang, Sebastian S Feger, Lucas Dullenkopf, Rulu Liao, Lukas Süsslin, Yuanting Liu, and Andreas Butz. Beyond recommendations: From backward to forward ai support of pilots' decision-making process. *arXiv preprint arXiv:2406.08959*, 2024.

[53] Mingqian Zheng, Jiaxin Pei, and David Jurgens. Is" a helpful assistant" the best role for large language models? a systematic evaluation of social roles in system prompts. *arXiv preprint arXiv:2311.10054*, 2023.