AI Systematically Rewires the Flow of Ideas

Anonymous authors
Paper under double-blind review

Abstract

As an epistemic technology, AI exerts influence over the ideas of individuals and society by acting as producer, mediator, and receiver of information. It impacts our collective knowledge, beliefs, and morality. In this position paper, we argue that there are mechanisms of hidden influence in AI development pipelines, which, when amplified by societal dynamics, could lead to dangerous outcomes that we may reverse by early interventions. We detail those mechanisms, amplifiers, and potential consequences in this paper.

1 Introduction

1.1 AI As an Epistemic Technology

AI has always been thought of as labor-replacement technology, automatic decision makers, or daily assistants. We argue that AI should be also regarded as *epistemic technology*, where AI is an active participant that shapes how do we perceive and understand our surroundings.

Defining Epistemic Technology Epistemic technologies are tools intentionally designed for acquiring, creating, manipulating, and disseminating knowledge. When studying these technologies, researchers primarily focus on epistemic concerns: specifically, how these tools modify epistemic contexts, content, and operations in ways that permanently alter both individual and collective approaches to perceiving and understanding the world.

As epistemic technology, AI can exert "epistemic" influence in humans' collective truth-seeking and morality forming process. By epistemic influence, we mean the influence is mainly onto the production of knowledge, such as microscopic, slide presentations, search engine, GPS navigation systems, and mathematical proofs, as opposed to production technology, which is mainly to enable speed and efficiency, such as hammers, pharmaceuticals, bulldozers, robotic arms, and IT systems (Alvarado, 2023). The point of framing this as an epistemic technology is we want to call attention to its potential epistemic harm onto humans, such as diversity loss (Padmakumar & He, 2023), and knowledge collapse in the longrun (Peterson, 2024) which are much less tangible, but their potential harms are no lesser significant.

1.2 Overview of AI Influence

Our position is that: AI exerts influence over the ideas of individuals, whether as a producer (e.g., content generator), mediator (e.g., recommender system), or receiver (e.g., preference learning from human feedback) of information. Such influence, which we term *AI influence*, can either be beneficial or harmful.

Empirical research on AI influence is ongoing but scattered. Those efforts are either clustered around specific affected subjects — Wikipedia (Wagner & Jiang, 2025), Stack Exchange community (Burtch et al., 2024), open-source community (Yeverechyahu et al., 2024), scientific publication and peer review (Liang et al., 2024a;b), political campaigns and elections (Hackenburg & Margetts, 2024; Potter et al., 2024) — or carved up along discipline boundaries like machine learning, cognitive science, and epistemology, with little cross-disciplinary discourse taking place.

AI influence is not necessarily a harm. Despite that we wanted to raise concerns around AI influence on human epistemology, it is too early to conclude that it is a bad thing. Humans are bound by

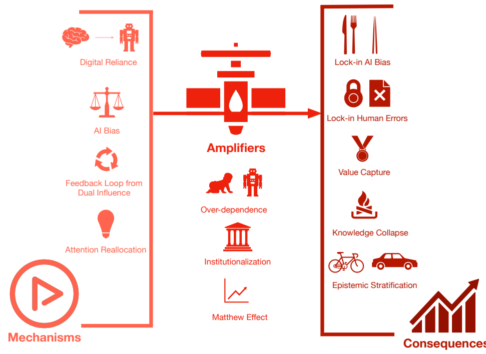

Figure 1: AI influence: mechanisms, amplifiers, and consequences.

cognitive limitations and it's likely that AI may expand our cognitive capacity and improve our collective deliberation.

1.3 Our Contributions

AI influence has deep roots in both AI safety research, where epistemic disruption has been identified as a key cause of catastrophic risks (Hendrycks & Mazeika, 2022), and AI ethics research, where the entrenchment of discriminatory biases has been a leading concern (Gross, 2023).

At the same time, AI influence has an explicit focus on the impact of AI via cognitive influence, as opposed to autonomous decision-making. Such a distinction implies differences in both research goals and methodology, both of which we highlight in this paper.

By introducing the framework of AI influence, we aim to synthesize fragmented research efforts and facilitate collaboration on this problem. The framework has the following notable traits.

Multi-agent and Human-AI interaction We think the problem of AI influence intrinsically focuses on multi-agent and human-AI interaction scenarios.

Knowledge, Beliefs, and Morality Our focus is AI's impact on humans' knowledge, beliefs, and morality, namely, what and how collectively humans perceive, understand, and value.

Long-term, Systematic, and Subtle We focus on long-term, systematic (including knock-on effects), and subtle influence from LLMs on human knowledge, beliefs, and morality. Those problems are often elusive and hard to evaluate. The methodologies to study and mitigate the problems are accordingly different from existing problems, notably those categorized as AI ethics and AI safety.

Mechanisms, Amplifiers, and Consequences We decompose the landscape of AI influence into three dimensions: *mechanisms*, the basic channels through which AI influences human epistemology; *amplifiers*, external factors that significantly enlarge such influence; and *consequences*, societal hazards that the amplified influence have led to or may soon lead to. These are *orthogonal* dimensions often connected by fully interconnected relationships, which allows us to focus on each of them individually while also touching on their connections.

Methodology to Study AI Influence We summarize all the methods that are used so far to study AI influence in Table 1. We also reason from first principles what could be used to study the influence from technologies on human truth-seeking and morality forming process. We illuminate the gap and discuss future research directions.

2 Mechanisms

In this section we cover specific mechanisms through which AI systems play a role in influencing human epistemics and morality, at an individual level and population level. By "mechanisms", we

Table 1: Classification of related research by methodology and topic.

		Qualitative Research	Formal Models	Simulations	Descriptive Analysis	Causal Inference	RCTs
Mechanisms	Digital Reliance	Hirvonen et al. (2024); Glickman & Sharot (2024a)			Burtch et al. (2024); Nirman et al. (2024); Thompson et al. (2024); Wagner & Jiang (2025)	Burtch et al. (2024); Gerlich (2025)	
	Human-AI Dual Influence	Brady et al. (2023); Brady & Crockett (2024); Collins et al. (2024); Lazar et al. (2024); Li & Yin (2024)	Lin et al. (2024); Ferbach et al. (2024); Krueger et al. (2020)	Wang et al. (2024); Brinkmann et al. (2022); Ferraro et al. (2024); Mansoury et al. (2020); Perra & Rocha (2019)	Li et al. (2023); Liang et al. (2024a)		Glickman & Sharot (2024b); Brinkmann et al. (2022); Chan et al. (2024); Pataranutaporn et al. (2023); Haupt et al. (2023); Hosseinmardi et al. (2024); Lu et al. (2024); Pappalardo et al. (2024); Sharma et al. (2024)
	Distinct AI Biases	Brandtzaeg et al. (2024); Köbis et al. (2021)	Taori & Hashimoto (2022)		Adilazuarda et al. (2024); Barman et al. (2024); Lamparth et al. (2024); Ryan et al. (2024)		Glickman & Sharot (2024b); Fisher et al. (2024); Danry et al. (2024); Costello et al. (2024); Kidd & Birhane (2023b); Kruegel et al. (2025); Leib et al. (2021); Piccardi et al. (2024); Potter et al. (2024)
	Attention Reallocation						Mendler-Dünner et al. (2024b); Haupt et al. (2023); Hosseinmardi et al. (2024)
Amplifiers	Trust				Araujo et al. (2020); Helberger et al. (2020)		Narayanan et al. (2023); Pataranutaporn et al. (2023); Reis et al. (2024)
	Institutional Path Dependence	Simon & Isaza-Ibarra (2023); Aoki (2024); Gruetzemacher et al. (2024); Lazar & Manuali (2024); Matz et al. (2024); Ovadya et al. (2024)					Potter et al. (2024)
	Socio-Economic Matthew Effect	Capraro et al. (2024)					
Consequences	Lock-in of Human Errors		Lin et al. (2024)	Wang et al. (2024); Mansoury et al. (2020); Perra & Rocha (2019)			Chan et al. (2024); Costello et al. (2024); Haupt et al. (2023); Kubin & Sikorski (2021)
	Lock-in of AI Biases	Brandtzaeg et al. (2024); Köbis et al. (2021)	Taori & Hashimoto (2022)		Adilazuarda et al. (2024); Barman et al. (2024); Lamparth et al. (2024); Ryan et al. (2024)		Glickman & Sharot (2024b); Fisher et al. (2024); Danry et al. (2024); Costello et al. (2024); Kidd & Birhane (2023b); Kruegel et al. (2025); Leib et al. (2021); Piccardi et al. (2024); Potter et al. (2024)
	Value Capture	Nguyen (2024b)					
	Knowledge Collapse	Brandtzaeg et al. (2024); Glickman & Sharot (2024a); Koskinen (2024); Wihbey (2024)		Peterson (2024); Bossens et al. (2024)	Burtch et al. (2024); Dohmatob et al. (2024); Thompson et al. (2024); Li et al. (2023); Liang et al. (2024a); Si et al. (2024); Wagner & Jiang (2025); Wu et al. (2024)	Anderson et al. (2024)	Doshi & Hauser (2023); Padmakumar & He (2023); Sharma et al. (2024)
	Epistemic Stratification	Kay et al. (2024)					

refer to either technical limitations of AI systems or new ways through which humans interact with AI may become sources of concerns over their epistemic impact over the long-term.

Here, we emphasize that the AI systems changes how information is originated, disseminated, propagated, and received, by humans or AI systems. The scope extends beyond that of algorithmic biases.

2.1 AI Introduces Distinct Biases into Collective Knowledge

Although AI systems are trained on data generated by humans, they do acquire distinctive biases from humans (Glickman & Sharot, 2024b; Kahneman et al., 2021). Specifically, there are the following reasons that introduce distinctive AI biases:

- *Token frequency more strongly impacts LLMs' output style than human's:* Tokens that are over-representative in dataset (e.g., words such as "significantly", "intricate", or "delve" appear dramatically more in academic writings in recent years (Geng et al., 2024; Liang et al., 2024a; Kobak et al., 2024);
- *LLMs are struggling with long-tail knowledge:* The accuracy of QnA correlates strongly with how many times questions and answers co-occur in the training dataset. (Kandpal et al., 2023; Das et al., 2024). On the other hand, LLMs rely on Retrieval-Augmented Generation (Lewis et al., 2020) to address LLMs' limitations dealing with long-tail knowledge. This approach, however, suffers from wasting compute resources on common knowledge since it indiscriminately retrieves documents.
- *Architectures create unique AI biases:* Architectural biases often stem from technical limitations, as opposed to biases in datasets that can be more readily resolved by more training or more data. One notable example is Convolutional Neural Networks (CNNs) are biased towards texture (Geirhos et al., 2018). Tokenization, the strategy LLM employs to split words into subwords, introduces biases unique to AI, such as downgrading arithmetic performances (Singh & Strouse, 2024), mishandling grammatical structures, and biases handling rare words (Phan et al., 2024).

Through training and deploying AI systems that acquire distinct biases, we risk introducing new biases the collective knowledge-making process (such as publication, journalism, scientific research etc). Such AI-biases might be persistent or even amplified because of digital reliance or feedback loops, as we will discuss in the following two subsections.

2.2 Cognitive Offloading, Cognitive Enhancements, and Digital Reliance

AI can enhance the human cognitive performance, which can take place either directly by providing advice and implementable solutions Senior et al. (2020); Fawzi et al. (2022) or indirectly by revealing novel cognitive strategies and problem-solving approaches Shin et al. (2023). Cognitive offloading is the term commonly used to describe such activities, namely, physical actions (such as preparing for a grocery list) to reduce cognitive demands required (Risko & Gilbert, 2016). Research shows that humans are willing to offload attention-demanding tasks to AI systems (Wahn et al., 2023). AI systems are also used to improve human cognitive performances. For example, A study that examines the performance of Go players Shin et al. (2023) reveals the performance of Go players improved after being exposed to AlphaGo moves, possibly as a result of learning novel non-human strategies from AlphaGo. Consistent results come from a study examining human problem-solving in a navigation task Brinkmann et al. (2022) . In this study, participants navigated through complex networks. Each path was associated with rewards (earning points) or penalties (losing points). Before performing the task, participants were exposed to solutions generated by the AI or by humans. The results demonstrated enhanced performance (accumulation of higher rewards) among players learning from AI, mainly due to the exposure to counterintuitive but optimal strategies generated by the AI. For example, the AI better identified than humans paths that initially appeared suboptimal but ultimately yielded better outcomes.

On the other hand, those cognitive offloading and enhancement activities enabled by AI may lead to digital reliance. Research demonstrates that reliance on digital tools, and in particular AI, alters different cognitive processes such as memory, critical thinking and problem-solving. For example, Sparrow et al. (2011) showed that when information is accessible through search engines, individuals prioritize remembering where to find this information rather than retaining it. This pattern extends to modern AI systems as well. Gerlich (2025) found that cognitive offloading to AI tools correlates with reduced critical thinking engagement, particularly among younger users who exhibit higher dependency. Consistent with these empirical findings, Zhai et al. (2024) conducted a systematic review revealing that over-reliance on AI dialogue systems impairs critical thinking and decision-making by fostering cognitive shortcuts. Together, these studies suggest that in some cases, delegating cognitive tasks to AI systems may deteriorate fundamental cognitive and thinking capabilities.

In this context of this position paper, digital reliance makes space for bias amplification, as we will discuss in the following subsections.

2.3 Human-AI Dual Influence Creates Feedback Loops

The influence between AI and humans is not one-directional. Humans' preferences can be influenced by the content generated by AI systems, while AI systems are trained to align with human preferences as well (e.g., Reinforcement Learning with Human Feedback Ziegler et al. (2019)). Such a feedback loop between humans and AI is similar to the feedback loop between content users and content creators in recommender systems, where users' tastes are shaped by the content they consume and creators produce content to fit users' tastes Jiang et al. (2019); Lin et al. (2024).

Although the human-AI dual influence mechanism might help to improve the alignment between humans and AI, it could also bring potential harms. For example, when humans or AI have initial biases or errors regarding a certain topic, such biases and errors can be circulated and amplified in human-AI interactions. There has been extensive research on human-to-AI and AI-to-human influence, but it was not until very recently research showed that the human-AI dual interaction may further exacerbate this influence mechanism: biased AI systems can affect human beliefs, rendering humans more biased compared to the initial state, due to the amplification of bias by AI systems and assigned trust by humans in AI judgements (Glickman & Sharot, 2024b;a).

AI bias is an established research field Mayson (2018). In this piece, however, we argue digital reliance on AI and feedback loops established in human-AI interactions legitimatize larger concerns over this topic. Not only because bias affects accuracy of medical decisions (Challen et al., 2019) or racial fairness (Salinas et al., 2023), which are by themselves important problems, but also because those biases are permanently introduced into epistemic processes and alter our worldviews (Vicente & Matute, 2023).

2.4 AI Reallocates Human Attention

One of the major functionalities of AI systems is they reorganize and redistribute information available to us, as search engines (including LLM-based ones) and RecSys-based social media do. In the previous subsection we cover new mechanism through which AI biases affect human judgements, while in this case, AI influence what do we see and think by selecting what information gets presented to us and receives our attention. This may have a strong agenda-setting effect on our thinking (Mendler-Dünner et al., 2024a).

It is worth getting very specific about the problem of attention and segmenting users with regard to potential problem. For sophisticated users of AI technologies, it is possible for generative models to be hugely creative, adding to intellectual diversity (Meincke et al., 2024). But such possibilities require careful technique and strategy, from few-shot prompting to chain of reasoning to iterative strategies generally. For the vast majority of the model-using public – which may not understand what the models are and do, and have little to little ability to execute prompt engineering strategies – usage may be largely passive and simplistic. Models will therefore tend to provide answers and content to the majority of users that conform to mainstream, modal patterns – the most likely next token, the probabilistic best answer or idea. This, in fact, is their central tendency and what they are architected to do. Using the models in a simplistic autocomplete or recommendation engine-style is likely to direct human attention to mainstream ideas and trends that are featured prominently on the open web (where the model pre-training has taken place), and not necessarily to more diverse, challenging, obscure, or marginal ideas or points.

3 Amplifiers

Mechanisms enumerated in §2 explain the forces that AI systems exert on human cognition and epistemology. These forces tend to be subtle and may not pose extreme risks on their own.

In this section, however, we introduce a range of *amplifiers* that are external to AI systems and may significantly increase AI influence, to the point of posing systemic risks described in §4.

3.1 Trust Amplifies AI Impact

Do higher levels of trust in AI correlate with increased AI influence? Recent research provides evidence supporting this claim. For example, Vicente and Matute Vicente & Matute (2023) demonstrated that higher trust in AI systems in medical diagnostic tasks led participants to adopt more of AI's biased recommendations, and even carrying these into subsequent tasks. Similarly, it was found that self-reported trust in AI systems was associated with the persuasiveness of deceptive AI classifications Danry et al. (2024). Interestingly, trust was not associated with additional the effect of additional AI-generated explanations.

Current evidence suggests that human trust in AI is highly sensitive to context and culture. While in many contexts, people prefer AI advice over humans' Araujo et al. (2020); Logg et al. (2019), in high stakes contexts (such as medicine or other life-threatening cases), people assign trust to humans more than they do to AI systems Reis et al. (2024). Additionally, Globig et al. (2024) found that trust in AI varies significantly across cultures Globig et al. (2024). Individuals in Eastern countries (e.g., India, Indonesia) exhibit greater trust and optimism towards AI compared to their Western counterparts (e.g., U.S., Germany), who tend to be more skeptical and cautious (Globig et al., 2024).

3.2 Institutional Path Dependence

Institutional path dependence refers to the tendency of organizations and systems to make decisions and adopt practices based on past trajectories, often locking in early patterns of behavior (Page et al., 2006). Epistemic frameworks through which institutions understand and address issues can be influenced by AI, an influence that can be hard to remove given the self-reinforcing nature of institutions (Arthur, 2018).

For instance, widespread AI application in the education sector may plant deep-rooted AI influence in children (Xu & Ouyang, 2022), AI advisors and analytics may bias governmental decision-making processes toward specific data-driven perspectives (Castelnovo & Sorrentino, 2021), AI-influenced public opinion can reinforce or challenge institutional norms (Panait & Ashraf, 2021), and early critical attitudes toward AI-generated art and writing has led to the enactment of institutional policies against the use of language models (Takagi, 2023; Kreitmeir & Raschky, 2023). Once these AI-mediated epistemic influences take root, their self-reinforcing nature may make it difficult to shift away from initial decisions, even in light of new evidence or changing contexts.

The self-reinforcing nature of the institutional path dependence problem will be particularly difficult to mitigate given recursion (Peterson, 2024). Once embedded narratives take hold and the climate of human opinion gets expressed at scale on social media and the web, AI models themselves will subsequently be trained on this new data containing AI influence. This "data coil" means path dependence becomes difficult to resist or reverse (Beer, 2022).

3.3 Socio-economic Matthew Effect

Advanced AI systems threatens to dramatically amplify existing socio-economic inequalities through what we term the "AI Matthew Effect," whereby initial advantages in AI access and capability compound exponentially over time.

Namely, it occurs when groups initially receiving more benefit from AI (*e.g.*, the wealthy, speakers of majority languages, those living in developed nations, those with access to GPUs, those working in fields where training data is more abundant) receive cascading benefits, and vice versa. An example is when biases against minority languages in LLMs shrink their user base who speaks minority languages, which could further reinforce biases against minority languages due to under-representation.

This dynamic could manifest through several interconnected mechanisms:

Productivity amplification: AI systems act as force multipliers for human productivity, with their effectiveness scaling in proportion to the user's existing capabilities and resources. High-skilled knowledge workers with access to state-of-the-art AI tools can leverage them to augment their expertise, potentially increasing their productivity by orders of magnitude. Meanwhile, workers in lower-skilled positions may find their jobs automated or devalued, widening the productivity gap.

Capital concentration: Organizations with early access to powerful AI systems can optimize operations, reduce costs, and capture market share more effectively than competitors. This advantage creates a self-reinforcing cycle where increased profits enable further AI investment and development, leading to market concentration.

4 Consequences

Influence mechanisms (§2), whose effects are magnified by amplifiers (§3), may lead to long-term consequences that are associated with large-scale hazards.

Long-term consequences are hard to clearly demonstrate in advance, but some have nonetheless manifested in empirical studies. Here we make a non-exhaustive list of these potential consequences.

AI systems that are trained on human data contains human errors and biases (such as cognitive biases (Binz & Schulz, 2023; Yax et al., 2024) can inherit the same biases and errors (Mayson, 2018). Interactions with those models can circulate those biases back to humans (Morewedge et al., 2023), even if humans do not have direct interactions with models (Valyaeva, 2024). Furthermore, those human errors and biases can be amplified via human-AI interactions because humans may assign more trust in AI output than average humans (Logg et al., 2019). The in-place mechanisms such as the psychological traits of humans and the training methods of LLMs raise concerns that those human errors and biases might be permanently preserved, amplified, and even locked into human society over the long run. The term Lock-in refers to the cases where values, beliefs, and knowledges memes/practices introduced into human society that last a long time, spread widely, assume a dominant position in ideology among a population, are are institutionalized (therefore hard to be removed), and cause damage (Hendrycks & Mazeika, 2022).

4.1 Lock-in of AI Biases

AI bias has been well documented and studied (Caliskan et al., 2017; Bolukbasi et al., 2016). However, a consequential effect has been largely overlooked: when humans interact with these biased systems, they internalize the systems' amplified bias becoming more biased than they initially were Glickman & Sharot (2024b); Vicente & Matute (2023). This bias amplification feedback loop relies on two key characteristics of AI systems: First, AI systems provide a higher signal-to-noise ratio compared to humans, consistently producing less variable outputs than human judgments (Kahneman et al., 2021). Second, in many domains, humans perceive AI systems as more capable and accurate than other humans Logg et al. (2019), making them more receptive to AI influence, or uncritically adopting AI biases. For instance, clinicians inherit AI biases even after AI systems are removed (Vicente & Matute, 2023). These characteristics create a dynamic where even small initial biases can be rapidly adopted and magnified through human-AI interactions. Furthermore This effect raises particular concerns for children, who have more malleable knowledge structures and may be more susceptible to AI's influence than adults (Kidd & Birhane, 2023a), making it concerning that such AI biases would be locked-in over generations.

4.2 Value Capture

Human objectives are often operationalized into quantifiable metrics — for instance, research quality being quantified as citation counts, and idea quality being quantified as the number of retweets.

Value capture happens when one mistakes quantified proxies for their much richer terminal values, and exclusively optimizes for the former instead, thereby losing the ability of personal deliberation on their values (Nguyen, 2024a).

AI has already been used in such quantification of objectives, for example in social media (Anandhan et al., 2018); other similar uses of AI has also been proposed, including as arbiters for resolving human disagreement (Tessler et al., 2024) and human representatives for collective decision-making (Zhang et al., 2024). In all such cases, human actors may be incentivised, or are already incentivised (Lüders et al., 2022; Wolf et al., 2017), to optimize for the AI-defined objectives. If such optimization becomes the dominant concerns of human participants — which is plausible given that AI products are often designed to be game-like and addictive (De et al., 2025) — value capture may steer people's values and objectives away from an ideal deliberative choice.

4.3 Knowledge Collapse

Knowledge collapse (Peterson, 2024) is defined as *the progressive narrowing over time of the set of information available to humans, along with a concomitant narrowing in the perceived availability and utility of different sets of information.* It is hypothesized to manifest as a "mode collapse" of collective knowledge in the human community, where long-detail information is lost while mainstream information is strengthened.

Peterson (2024) mainly focuses on unrepresentative data, lack of in-depth exploration during LLM inference, and algorithmic limitations of next-token prediction as the potential causes of knowledge collapse. Peterson (2024) argues that by making mainstream information more readily available, LLMs shift attention away from long-tail information.

In addition to these concerns, we note that other mechanisms outlined in this paper, including for example dual influence (Lin et al., 2024), can similarly contribute to knowledge collapse. From a mechanistic angle, knowledge collapse and lock-in share many commonalities, most especially the reinforcement of existing popular ideas and the suppression of marginal ones.

4.4 Epistemic Stratification

Epistemic stratification is the unequal distribution of access to knowledge, resources, and cognitive tools across individuals or groups, leading to disparities in their ability to acquire, evaluate, and generate knowledge (Silva Filho et al., 2023).

AI may contribute to epistemic stratification by amplifying existing disparities, such as through unequal access to advanced AI tools, biased algorithmic recommendations that reinforce echo chambers, the prioritization of information access for privileged demographics (Kay et al., 2024), or the increasingly centralized control over AI development (Brynjolfsson & Ng, 2023).

5 Alternative Views

5.1 AI Systems Have Negligible Impact on Human Cognition

Empirical evidence does not provide with a holistic picture of AI's impact on human cognition. It is true that humans are becoming more reliant on AI systems for their tasks, but it is unclear whether having AI systems to process those tasks for humans would necessarily degenerate or enhance those cognitive capabilities of humans. At least, it is still unclear whether humans' navigation skills are compromised because of using GPS tools (Fricker, 2021; Jadallah et al., 2017).

Two questions that are instrumental to understand AI systems' impact on human cognition. For one, does digital reliance influence human cognitive skills that are directly replaced by corresponding AI capabilities (Teschke et al., 2013)? For instance, does the use of GPS hurt human navigation skills (Fricker, 2021)? The same question could be asked about other digital tools and human skills, such as calculators and arithmetic skills, machine translation tools and second language acquisition skills. For the other, do the replaced domain-specific human skills undermine more general human cognitive capabilities? For example, does the undermined arithmetic skills hurt human general mathematical reasoning and problem-solving abilities (Geary et al., 2015; Hurst & Cordes, 2018)?

Without sufficient empirical evidence on how might human cognition be altered in the presence of new tools, especially AI systems, it is hard to firmly hold our position. Hence, an alternative view is, AI systems may have negligible impact on human cognitive capabilities over the long-term. One reason this might be true is we do not understand the relationship between low-level domain-specific skills to high-level general capabilities. Replacing the former by AI may have little negative impact on the latter, or using tool may enhance cognitive capabilities (Teschke et al., 2013).

5.2 Highly Parameterized AI Systems Are Less Biased and Error-prone than Humans Are

AI systems are biased (Jadallah et al., 2017) and error-prone (Zhou et al., 2024), as research has revealed, but so are humans. Besides those inductive biases that are introduced by specific architectures and training methods (Geirhos et al., 2018; Singh & Strouse, 2024), AI systems acquire their biases from training datasets and by extension, from humans. Highly parameterized AI systems such as LLMs are less biased and error-prone than conventional machine learning models because LLMs are more expressive and techniques such as RAG helps to consult external sources for truth validation (Gao et al., 2023). Meanwhile, it is also likely that state-of-arts AI systems may become even less biased and error-prone than average humans are. From the point of view of collective truth-seeking (such as conducting scientific research and collective deliberation), AI systems functioning as ”shadow authors” to individual humans can be positive.

That being said, err should be on the side of being cautious. It is likely that AI biases, errors, and hallucinations become more elusive before they are removed (Zhou et al., 2024). Once they are hard to be found by average users, commercial developers are much motivated to address those problems, creating persistent and even amplified biases and errors (Ren et al., 2024), which are precisely what we warn in this piece.

5.3 AI’s Epistemic Impact Can be Positive

In 4 we have detailed AI’s long-term impact on human knowledge and values. Notably, they seem overwhelmingly negative. As much as we strive to be epistemic neural on AI influence, we are biased and limited in our perspectives. It is not our intention to present negative views only. We want to raise attention on AI’s epistemic over the long-term and avoid the knock-on effects over the long-term, but we also want to acknowledge we are far from having a holistic picture.

It is entirely likely the issues we have raised here can be addressed over time and people can become wise in using those tools. For instance, users, especially students and researchers may acquire a critical lens of AI’s generated content. Such a field is called ”AI literacy”, which is to teach individuals to use, understand, and evaluate AI systems (Casal-Otero et al., 2023). Sufficient critical thinking skills, paired with AI systems increasing reach and capabilities may cultivate a generation of more informed learners and citizens who are more capable to participate in collective truth-seeking and deliberative processes.

6 Conclusion

AI exerts systematic influence over the ideas in individuals and society. We have outlined the mechanisms that enable such influence, the amplifiers that magnify the influence, and potential consequences it may entail.

The eventual aim of AI influence research is to enable the responsible management of AI influence, reaping its benefits while avoiding the harms. It is crucial in light of the rapid development of generative models, which exerts large influence on human users and society that is not seen before.

Accomplishing such an aim requires coordination between AI safety and ethics communities, machine learning and human-computer interaction communities, social science communities, and importantly, industry actors. We hope this paper could be the start of a partnership between different communities in search for robust solutions for the management of AI influence.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. Towards

Measuring and Modeling ”Culture” in LLMs: A Survey. *arXiv*, 2024. doi: 10.48550/arxiv.2403.15412.

Ramón Alvarado. AI as an Epistemic Technology. *Science and Engineering Ethics*, 29(5):32, 2023. ISSN 1353-3452. doi: 10.1007/s11948-023-00451-3.

Anitha Anandhan, Liyana Shuib, Maizatul Akmar Ismail, and Ghulam Mujtaba. Social media recommender systems: review and open research issues. *IEEE Access*, 6:15608–15628, 2018.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization Effects of Large Language Models on Human Creative Ideation. *Creativity and Cognition*, pp. 413–425, 2024. doi: 10.1145/3635636.3656204.

Goshi Aoki. Large Language Models in Politics and Democracy: A Comprehensive Survey. *arXiv*, 2024. doi: 10.48550/arxiv.2412.04498.

Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, 35(3): 611–623, 2020. ISSN 0951-5666. doi: 10.1007/s00146-019-00931-w.

W Brian Arthur. Self-reinforcing mechanisms in economics. In *The economy as an evolving complex system*, pp. 9–31. CRC Press, 2018.

Kristian González Barman, Simon Lohse, and Henk de Regt. Reinforcement Learning from Human Feedback: Whose Culture, Whose Values, Whose Perspectives? *arXiv*, 2024.

David Beer. The problem of researching a recursive society: Algorithms, data coils and the looping of the social. *Big Data & Society*, 9(2):20539517221104997, 2022.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

David M Bossens, Shanshan Feng, and Yew-Soon Ong. The Digital Ecosystem of Beliefs: does evolution favour AI over humans? *arXiv*, 2024. doi: 10.48550/arxiv.2412.14500.

William J. Brady and M. J. Crockett. Norm Psychology in the Digital Age: How Social Media Shapes the Cultural Evolution of Normativity. *Perspectives on Psychological Science*, 19(1):62–64, 2024. ISSN 1745-6916. doi: 10.1177/17456916231187395.

William J. Brady, Joshua Conrad Jackson, Björn Lindström, and M.J. Crockett. Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, 27(10):947–960, 2023. ISSN 1364-6613. doi: 10.1016/j.tics.2023.06.008.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. Understanding model power in social AI. *AI & SOCIETY*, pp. 1–11, 2024. ISSN 0951-5666. doi: 10.1007/s00146-024-02053-4. Table 1 contains a collection of HCI studies.

L. Brinkmann, D. Gezerli, K. V. Kleist, T. F. Mller, I. Rahwan, and N. Pescetelli. Hybrid social learning in human-algorithm cultural transmission. *Philosophical Transactions of the Royal Society A*, 380(2227):20200426, 2022. ISSN 1364-503X. doi: 10.1098/rsta.2020.0426.

Erik Brynjolfsson and Andrew Ng. Big ai can centralize decision-making and power, and that'sa problem. *Missing links in AI governance*, pp. 65, 2023.

Gordon Burtch, Dokyun Lee, and Zhichen Chen. The consequences of generative AI for online knowledge communities. *Scientific Reports*, 14(1):10413, 2024. doi: 10.1038/s41598-024-61221-0.

Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas, Jim A C Everett, Gerd Gigerenzer, Christine Greenhow, Daniel A Hashimoto, Julianne Holt-Lunstad, Jolanda Jetten, Simon Johnson, Werner H Kunz, Chiara Longoni, Pete Lunn, Simone Natale, Stefanie Paluch, Iyad Rahwan, Neil Selwyn, Vivek Singh, Siddharth Suri, Jennifer Sutcliffe, Joe Tomlinson, Sander van der Linden, Paul A M Van Lange, Friederike Wall, Jay J Van Bavel, and Riccardo Viale. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6):pgae191, 2024. doi: 10.1093/pnasnexus/pgae191.

Lorena Casal-Otero, Alejandro Catala, Carmen Fernández-Morante, Maria Taboada, Beatriz Cebreiro, and Senén Barro. Ai literacy in k-12: a systematic literature review. *International Journal of STEM Education*, 10(1):29, 2023.

Walter Castelnovo and Maddalena Sorrentino. The nodality disconnect of data-driven government. *Administration & Society*, 53(9):1418–1442, 2021.

Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ quality & safety*, 28(3):231–237, 2019.

Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F Loftus. Conversational AI Powered by Large Language Models Amplifies False Memories in Witness Interviews. *arXiv*, 2024.

Katherine M. Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B. Tenenbaum, and Thomas L. Griffiths. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024. doi: 10.1038/s41562-024-01991-9.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814, 2024. ISSN 0036-8075. doi: 10.1126/science.adq1814.

Valdemar Danry, Pat Pataranutaporn, Matthew Groh, Ziv Epstein, and Pattie Maes. Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation. *arXiv*, 2024.

Debarati Das, Karin De Langis, Anna Martin-Boyle, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Anugrah Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*, 2024.

Debasmita De, Mazen El Jamal, Eda Aydemir, and Anika Khera. Social media algorithms and teen addiction: Neurophysiological impact and ethical considerations. *Cureus*, 17(1), 2025.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A Tale of Tails: Model Collapse as a Change of Scaling Laws. *arXiv*, 2024. doi: 10.48550/arxiv.2402.07043.

Anil R Doshi and Oliver P Hauser. Generative artificial intelligence enhances creativity but reduces the diversity of novel content. *arXiv*, 2023. doi: 10.48550/arxiv.2312.00506.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences. *arXiv*, 2024. doi: 10.48550/arxiv.2407.09499.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. Agent-Based Modelling Meets Generative AI in Social Network Simulations. *arXiv*, 2024. doi: 10.48550/arxiv.2411.16031.

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. Biased AI can Influence Political Decision-Making. *arXiv*, 2024.

Elizabeth Fricker. Should we worry about silicone chip technology de-skilling us? *Royal Institute of Philosophy Supplements*, 89:131–152, 2021.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

David C Geary, Mary K Hoard, Lara Nugent, and Jeffrey N Rouder. Individual differences in algebraic cognition: Relation to the approximate number and semantic memory systems. *Journal of experimental child psychology*, 140:211–227, 2015.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Mingmeng Geng, Caixi Chen, Yanru Wu, Dongping Chen, Yao Wan, and Pan Zhou. The impact of large language models in academia: from writing to speaking. *arXiv preprint arXiv:2409.13686*, 2024.

Michael Gerlich. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1):6, 2025. doi: 10.3390/soc15010006.

Moshe Glickman and Tali Sharot. AI-induced hyper-learning in humans. *Current Opinion in Psychology*, 60:101900, 2024a. ISSN 2352-250X. doi: 10.1016/j.copsyc.2024.101900.

Moshe Glickman and Tali Sharot. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, pp. 1–15, 2024b.

Laura K Globig, Rachel Xu, Steve Rathje, and Jay J Van Bavel. Perceived (mis) alignment in generative artificial intelligence varies across cultures. *Preprint. DOI*, 10, 2024.

Nicole Gross. What chatgpt tells us about gender: a cautionary tale about performativity and gender biases in ai. *Social Sciences*, 12(8):435, 2023.

Ross Gruetzemacher, Shahar Avin, James Fox, and Alexander K Saeri. Strategic Insights from Simulation Gaming of AI Race Dynamics. *arXiv*, 2024.

Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24): e2403116121, 2024.

Andreas Haupt, Mihaela Curmei, François-Marie de Jouvencel, Marc Faddoul, Benjamin Recht, and Dylan Hadfield-Menell. The long-term effects of personalization: Evidence from youtube. 2023.

Natali Helberger, Theo Araujo, and Claes H. de Vreese. Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, 39:105456, 2020. ISSN 0267-3649. doi: 10.1016/j.clsr.2020.105456.

Dan Hendrycks and Mantas Mazeika. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*, 2022.

Noora Hirvonen, Ville Jylhä, Yucong Lao, and Stefan Larsson. Artificial intelligence in the information ecosystem: Affordances for everyday information seeking. *Journal of the Association for Information Science and Technology*, 75(10):1152–1165, 2024. ISSN 2330-1635. doi: 10.1002/asi.24860.

Homa Hosseinmardi, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, and Duncan J. Watts. Causally estimating the effect of YouTube's recommender system using counterfactual bots. *Proceedings of the National Academy of Sciences*, 121(8):e2313377121, 2024. ISSN 0027-8424. doi: 10.1073/pnas.2313377121.

Michelle Hurst and Sara Cordes. A systematic investigation of the link between rational number processing and algebra ability. *British Journal of Psychology*, 109(1):99–117, 2018.

May Jadallah, Alycia M Hund, Jonathan Thayn, Joel Garth Studebaker, Zachary J Roman, and Elizabeth Kirby. Integrating geospatial technologies in fifth-grade curriculum: Impact on spatial ability and map-analysis skills. *Journal of Geography*, 116(4):139–151, 2017.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate Feedback Loops in Recommender Systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 383–390, Honolulu HI USA, January 2019. ACM. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314288. URL `https://dl.acm.org/doi/10.1145/3306618.3314288`.

Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. *Noise: A flaw in human judgment*. Hachette UK, 2021.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pp. 15696–15707. PMLR, 2023.

Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. Epistemic Injustice in Generative AI. *arXiv*, 2024. doi: 10.48550/arxiv.2408.11441.

Celeste Kidd and Abeba Birhane. How ai can distort human beliefs. *Science*, 380(6651):1222–1223, 2023a.

Celeste Kidd and Abeba Birhane. How AI can distort human beliefs. *Science*, 380(6651):1222–1223, 2023b. ISSN 0036-8075. doi: 10.1126/science.adi0248.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into chatgpt usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016*, 2024.

Nils Köbis, Jean-François Bonnefon, and Iyad Rahwan. Bad machines corrupt good morals. *Nature Human Behaviour*, 5(6):679–685, 2021. doi: 10.1038/s41562-021-01128-2.

Inkeri Koskinen. We Have No Satisfactory Social Epistemology of AI-Based Science. *Social Epistemology*, 38(4):458–475, 2024. ISSN 0269-1728. doi: 10.1080/02691728.2023.2286253.

David H Kreitmeir and Paul A Raschky. The unintended consequences of censoring digital technology–evidence from italy's chatgpt ban. *arXiv preprint arXiv:2304.09339*, 2023.

Sebastian Kruegel, Andreas Ostermaier, and Matthias Uhl. ChatGPT's advice drives moral judgments with or without justification. *arXiv*, 2025. doi: 10.48550/arxiv.2501.01897.

David Krueger, Tegan Maharaj, and Jan Leike. Hidden Incentives for Auto-Induced Distributional Shift. *arXiv*, 2020.

Emily Kubin and Christian von Sikorski. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206, 2021. ISSN 2380-8985. doi: 10.1080/23808985.2021.1976070.

Max Lamparth, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider, and Harold Trinkunas. Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations. *arXiv*, 2024. doi: 10.48550/arxiv.2403.03407.

Seth Lazar and Lorenzo Manuali. Can LLMs advance democratic values? *arXiv*, 2024.

Seth Lazar, Luke Thorburn, Tian Jin, and Luca Belli. The Moral Case for Using Language Model Agents for Recommendation. *arXiv*, 2024.

Margarita Leib, Nils C Köbis, Rainer Michael Rilke, Marloes Hagens, and Bernd Irlenbusch. The corruptive force of AI-generated advice. *arXiv*, 2021.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Chao Li, Xing Su, Haoying Han, Cong Xue, Chunmo Zheng, and Chao Fan. Quantifying the Impact of Large Language Models on Collective Opinion Dynamics. *arXiv*, 2023. doi: 10.48550/arxiv.2308.03313.

Zhuoyan Li and Ming Yin. Utilizing Human Behavior Modeling to Manipulate Explanations in AI-Assisted Decision Making: The Good, the Bad, and the Scary. *arXiv*, 2024. doi: 10.48550/arxiv.2411.10461.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y Zou. Mapping the Increasing Use of LLMs in Scientific Papers. *arXiv*, 2024a.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024b.

Tao Lin, Kun Jin, Andrew Estornell, Xiaoying Zhang, Yiling Chen, and Yang Liu. User-Creator Feature Dynamics in Recommender Systems with Dual Influence. *arXiv*, 2024. doi: 10.48550/arxiv.2407.14094.

Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151: 90–103, 2019.

Jinwei Lu, Yikuan Yan, Keman Huang, Ming Yin, and Fang Zhang. Do We Learn From Each Other: Understanding the Human-AI Co-Learning Process Embedded in Human-AI Collaboration. *Group Decision and Negotiation*, pp. 1–37, 2024. ISSN 0926-2644. doi: 10.1007/s10726-024-09912-x.

Adrian Lüders, Alejandro Dinkelberg, and Michael Quayle. Becoming "us" in digital spaces: How online users creatively and strategically exploit social media affordances to build up social identity. *Acta Psychologica*, 228:103643, 2022.

Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback Loop and Bias Amplification in Recommender Systems. *arXiv*, 2020. doi: 10.48550/arxiv.2007.13019.

S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1):4692, 2024. doi: 10.1038/s41598-024-53755-0.

Sandra G Mayson. Bias in, bias out. *YAle lJ*, 128:2218, 2018.

Lennart Meincke, Ethan R Mollick, and Christian Terwiesch. Prompting diverse ideas: Increasing ai idea variance. *arXiv preprint arXiv:2402.01727*, 2024.

Celestine Mendler-Dünner, Gabriele Carovano, and Moritz Hardt. An engine not a camera: Measuring performative power of online search. *arXiv preprint arXiv:2405.19073*, 2024a.

Celestine Mendler-Dünner, Gabriele Carovano, and Moritz Hardt. An engine not a camera: Measuring performative power of online search. *arXiv*, 2024b. doi: 10.48550/arxiv.2405.19073.

Carey K Morewedge, Sendhil Mullainathan, Haaya F Naushan, Cass R Sunstein, Jon Kleinberg, Manish Raghavan, and Jens O Ludwig. Human bias in algorithm design. *Nature Human Behaviour*, 7(11):1822–1824, 2023.

Saumik Narayanan, Guanghui Yu, Chien-Ju Ho, and Ming Yin. How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making? *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 49–57, 2023. doi: 10.1145/3600211.3604709.

Christopher Nguyen. Value capture. *Journal of Ethics and Social Philosophy*, 27(3), 2024a. doi: 10.26556/jesp.v27i3.3048.

Christopher Nguyen. Value Capture. *Journal of Ethics and Social Philosophy*, 27(3), 2024b. doi: 10.26556/jesp.v27i3.3048.

Diana Bar-Or Nirman, Ariel Weizman, and Amos Azaria. Fool Me, Fool Me: User Attitudes Toward LLM Falsehoods. *arXiv*, 2024. doi: 10.48550/arxiv.2412.11625.

Aviv Ovadya, Luke Thorburn, Kyle Redman, Flynn Devine, Smitha Milli, Manon Revel, Andrew Konya, and Atoosa Kasirzadeh. Toward Democracy Levels for AI. *arXiv*, 2024. doi: 10.48550/arxiv.2411.09222.

Vishakh Padmakumar and He He. Does Writing with Language Models Reduce Content Diversity? *arXiv*, 2023.

Scott E Page et al. Path dependence. *Quarterly Journal of Political Science*, 1(1):87–115, 2006.

Cezara Panait and Cameran Ashraf. Ai algorithms–(re) shaping public opinions through interfering with access to information in the online environment. *Europuls Policy Journal*, 1(1):46–64, 2021.

Luca Pappalardo, Emanuele Ferragina, Salvatore Citraro, Giuliano Cornacchia, Mirco Nanni, Giulio Rossetti, Gizem Gezici, Fosca Giannotti, Margherita Lalli, Daniele Gambetta, Giovanni Mauro, Virginia Morini, Valentina Pansanella, and Dino Pedreschi. A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions. *arXiv*, 2024.

Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10):1076–1086, 2023. doi: 10.1038/s42256-023-00720-7.

Nicola Perra and Luis E. C. Rocha. Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific Reports*, 9(1):7261, 2019. doi: 10.1038/s41598-019-43830-2.

Andrew J Peterson. AI and the Problem of Knowledge Collapse. *arXiv*, 2024.

Buu Phan, Marton Havasi, Matthew Muckley, and Karen Ullrich. Understanding and mitigating tokenization bias in language models. *arXiv preprint arXiv:2406.16829*, 2024.

Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey T Hancock, Jeanne L Tsai, and Michael Bernstein. Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity. *arXiv*, 2024.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. *arXiv*, 2024.

Moritz Reis, Florian Reis, and Wilfried Kunde. Influence of believed AI involvement on the perception of digital medical advice. *Nature Medicine*, 30(11):3098–3100, 2024. ISSN 1078-8956. doi: 10.1038/s41591-024-03180-7.

Yi Ren, Shangmin Guo, Linlu Qiu, Bailin Wang, and Danica J Sutherland. Language Model Evolution: An Iterated Learning Perspective. *arXiv*, 2024. doi: 10.48550/arxiv.2404.04286.

Evan F Risko and Sam J Gilbert. Cognitive offloading. *Trends in cognitive sciences*, 20(9):676–688, 2016.

Michael J Ryan, William Held, and Diyi Yang. Unintended Impacts of LLM Alignment on Global Representation. *arXiv*, 2024.

Abel Salinas, Louis Penafiel, Robert McCormack, and Fred Morstatter. " im not racist but...": Discovering bias in the internal knowledge of large language models. *arXiv preprint arXiv:2310.08780*, 2023.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024. doi: 10.1145/3613904.3642459.

Minkyu Shin, Jin Kim, Bas Van Opheusden, and Thomas L Griffiths. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12):e2214840120, 2023.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv*, 2024.

Waldomiro J Silva Filho, Maria Virginia M Dazzani, Luca Tateo, Rodrigo Gottschalk Sukerman Barreto, and Giuseppina Marsico. He knows, she doesn't? epistemic inequality in a developmental perspective. *Review of General Psychology*, 27(3):231–244, 2023.

Felix M Simon and Luisa Fernanda Isaza-Ibarra. Ai in the news: reshaping the information ecosystem? 2023.

Aaditya K Singh and DJ Strouse. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *arXiv preprint arXiv:2402.14903*, 2024.

Betsy Sparrow, Jenny Liu, and Daniel M Wegner. Google effects on memory: Cognitive consequences of having information at our fingertips. *science*, 333(6043):776–778, 2011.

Nicole Miu Takagi. Banning of chatgpt from educational spaces: A reddit perspective. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pp. 179–194, 2023.

Rohan Taori and Tatsunori B Hashimoto. Data Feedback Loops: Model-driven Amplification of Dataset Biases. *arXiv*, 2022.

Irmgard Teschke, Claudia AF Wascher, Madeleine F Scriba, Auguste MP von Bayern, V Huml, B Siemers, and Sabine Tebbich. Did tool-use evolve with enhanced physical cognitive abilities? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1630):20120418, 2013.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024. ISSN 0036-8075. doi: 10.1126/science.adq2852.

Brian Thompson, Mehak Preet Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism. *arXiv*, 2024.

A. Valyaeva. Ai has already created as many images as photographers have taken in 150 years. *Everypixel Journal*, 2024. URL `https://journal.everypixel.com/ai-image-statistics`. Accessed: [Insert Date].

Lucía Vicente and Helena Matute. Humans inherit artificial intelligence biases. *Scientific Reports*, 13 (1):15737, 2023.

Christian Wagner and Ling Jiang. Death by AI: Will large language models diminish Wikipedia? *Journal of the Association for Information Science and Technology*, 2025. ISSN 2330-1635. doi: 10.1002/asi.24975.

Basil Wahn, Laura Schmitz, Frauke Nora Gerster, and Matthias Weiss. Offloading under cognitive load: Humans are willing to offload parts of an attentionally demanding task to an algorithm. *Plos one*, 18(5):e0286102, 2023.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. Decoding Echo Chambers: LLM-Powered Simulations Revealing Polarization in Social Networks. *arXiv*, 2024. doi: 10.48550/arxiv.2409.19338.

John Wihbey. AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge? *SSRN Electronic Journal*, 2024. doi: 10.2139/ssrn.4805026.

Marty J Wolf, K Miller, and Frances S Grodzinsky. Why we should have seen that coming: comments on microsoft's tay" experiment," and wider implications. *Acm Sigcas Computers and Society*, 47 (3):54–64, 2017.

Fan Wu, Emily Black, and Varun Chandrasekaran. Generative Monoculture in Large Language Models. *arXiv*, 2024.

Weiqi Xu and Fan Ouyang. A systematic review of ai role in the educational system based on a proposed conceptual framework. *Education and Information Technologies*, 27(3):4195–4223, 2022.

Nicolas Yax, Hernan Anlló, and Stefano Palminteri. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51, 2024.

Doron Yeverechyahu, Raveesh Mayya, and Gal Oestreicher-Singer. The impact of large language models on open-source innovation: Evidence from github copilot. *arXiv preprint arXiv:2409.08379*, 2024.

Chunpeng Zhai, Santoso Wibowo, and Lily D Li. The effects of over-reliance on ai dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28, 2024.

Zhaowei Zhang, Fengshuo Bai, Mingzhi Wang, Haoyang Ye, Chengdong Ma, and Yaodong Yang. Incentive compatibility for ai alignment in sociotechnical systems: Positions and prospects. *arXiv preprint arXiv:2402.12907*, 2024.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.