POSITION: AI DEVELOPMENT SHOULD PRIORITIZE COGNITIVE SECURITY

Batu El[♣1], **Shiye Su**[♣2], **Aneesh Pappu**[♣3], **Peggy Yin**[♣4], **Julie Heng**[♣5], **Eric Heng**[6], **Ryan Wang**[7], **Andreas Haupt**[♠2 8 9], **James Zou**[♠2 3 10]

[1]Institute for Computational and Mathematical Engineering, Stanford University
[2]Department of Computer Science, Stanford University
[3]Department of Electrical Engineering, Stanford University
[4]Department of Psychology, Stanford University
[5]Stanford Law School, Stanford University
[6]Department of Biology, Stanford University
[7]Department of Bioengineering, Stanford University
[8]Stanford Institute for Human-Centered AI (HAI)
[9]Department of Economics, Stanford University
[10]Department of Biomedical Data Science, Stanford University

♣ Equal contribution ♠ Equal senior authorship

ABSTRACT

Generative AI systems designed to influence human beliefs and actions are becoming increasingly pervasive, raising concerns about *cognitive security*–the protection of human cognitive processes from hazardous influence. Recent advances have only amplified the cognitive risks of AI technologies, leading institutions worldwide to identify cognitive security as an emerging and urgent governance challenge. However, research on such cognitive impacts remains ad hoc and fragmented across the literature, and the field lacks a framework to address them. In this paper, we argue that generative AI research and development should prioritize cognitive security. First, we track and categorize state-of-the-art capabilities for cognitive influence in generative AI systems. Then, we outline a roadmap for advancing cognitive security research within AI research by (1) proposing attack and defense mechanisms to formalize threat models, expose vulnerabilities, and evaluate countermeasures and (2) defining metrics to measure cognitive impacts.

1 INTRODUCTION

Human beliefs and actions are increasingly shaped by generative AI systems designed to observe, predict, and influence them. AI-generated content can now change political attitudes (Schoenegger et al., 2025; Matz et al., 2024; Lin et al., 2025; Hackenburg et al., 2025; Costello et al., 2024b),[1] increase user engagement with digital systems and content. (Coppolillo et al., 2025; Wang et al., 2025), and drive consumer behavior and purchasing decisions (Zhang et al., 2025; Zhu et al., 2025). Meanwhile, developments in adjacent areas, such as elicitation (Skeggs et al., 2025; Li et al., 2023; Ma et al., 2025), prediction of human behavior (Park et al., 2023; Aher et al., 2023; Argyle et al., 2023; Binz & Schulz, 2023; Abdulhai et al., 2025a), and alternative modalities (Slater & Sanchez-Vives, 2016; Bozkir et al., 2023; Ramirez et al., 2013; Musk & Neuralink, 2019),[2] have the potential to substantially amplify these influence capabilities. Alarmed by these developments, policymakers and policy researchers alike are warning that AI capabilities in generating media (Huang et al., 2023; Vigliarolo, 2023) and in personalization (Lange, 2024; Daniel Nikoula, 2024) have enabled operations on cognitive processes by "hostile states, political movements, and profit-driven platforms" (Bicakci,

[1]Persuasion references include general persuasive abilities (Schoenegger et al., 2025), personalized persuasive messaging (Matz et al., 2024), political persuasion (Lin et al., 2025; Hackenburg et al., 2025), and reducing beliefs in conspiracy theories (Costello et al., 2024b)

[2]References for alternative modalities include virtual worlds (Slater & Sanchez-Vives, 2016; Bozkir et al., 2023) and brain-computer interfaces (Ramirez et al., 2013; Musk & Neuralink, 2019).

2025; Catena et al., 2025b; Pamment & Tsurtsumia, 2025; Hung & Hung, 2022). *While advances in AI development have enabled these threats, the field lacks a framework for addressing the risks it creates.*

Cognitive Security. In this paper, we define cognitive security as the *protection of cognitive processes from hazardous influence*. By *cognitive processes*, we refer to the mechanisms through which an agent uses its accumulated dispositions to map its current experience to actions or expressions of beliefs and attitudes. Defining *hazardous influence* is difficult because any proposed boundary invokes contested normative commitments (Susser et al., 2019a;b; Coons & Weber, 2014; Jongepier & Klenk, 2022; Thaler & Sunstein, 2008; Carroll et al., 2023b). We therefore adopt a portable definition that supports reasoning about hazards across different normative frameworks. By *hazardous influence*, we refer to any act of influence that is exercised with intent to, or that carries a reasonable potential to: (a) induce or facilitate conduct that is unlawful; (b) cause harm to the person influenced; or (c) impair the integrity of the public institutions that rely on the faithful aggregation of individual preferences. Rather than prescribing specific values, our definition offers a framework whose substantive content (what is unlawful, what qualifies as harm, and which institutions are relevant) is supplied by the governing legal jurisdiction or normative framework (whether professional, religious, organizational, or cultural) that holds authority in the given context.

Previous definitions of cognitive security span a range of framings, from defending against psychosocial manipulation and information warfare (Waltzman, 2017; Catena et al., 2025a) to protecting against threats in the cognitive domain (Doherty, 2023; Crum et al., 2025b) to augmenting the situational awareness of security analysts through AI (Ask et al., 2025), and applying AI to detect and mitigate cybersecurity threats (Casino, 2025). Building on these prior definitions, our definition of cognitive security similarly focuses on protecting individuals and collectives from harmful influence targeting their cognitive processes, while offering a new portable framework that minimizes prescriptive value commitments and extending concern beyond individual cognition to the integrity of collective decision-making institutions. Our definition is agnostic to the source of cognitive threat, which may be states, corporations, or other actors.

We argue in this paper that the capabilities of generative AI give rise to especially acute and personalized techniques for eliciting from, predicting, and influencing humans. This motivates cognitive security as an AI research priority.

Research Agenda. We frame influence over human cognition as a security problem, analogous to adversarial attacks on machine learning systems. Building on this framing, we outline a systematic attack–defense research cycle to enable the development and evaluation of proactive and reactive defenses designed to preserve cognitive autonomy. Finally, we propose a suite of metrics to operationalize cognitive security, enabling rigorous measurement of belief change, durability, dependence, and awareness of AI influence.

Outline. In the following sections, we show how cognitive security underpins core functions at the individual and societal level (Section 2) and highlight how recent advances in generative AI amplify cognitive security risks (Section 3). This motivates our cognitive security research agenda (Section 4), in which we propose a framework for developing attack and defense mechanisms and define metrics to measure cognitive impacts. Finally, we present alternative viewpoints (Section 5).

2 Human Cognition as a Security-Critical Asset

Consider a simple model of cognitive processes: a human encounters inputs from the world and uses their accumulated dispositions to map their current experience to actions or expressions of beliefs and attitudes. Now, imagine *a system capable of reliably predicting which inputs will induce a specific behavior or expression in a human*. Such a system requires neither coercion nor deception to manipulate its users: it simply optimizes the sequence of content presented to the user to shape the user's beliefs and actions. In other words, beliefs and actions are no longer emerging from interactions with a neutral environment. Instead, they are engineered by an optimizing process that solves an inverse problem, selecting inputs with the goal of obtaining a particular output, in order to push individuals toward specific beliefs or actions, such as favoring a political candidate (Lin et al., 2025) or continuing to engage with a digital interface (Milli et al., 2025) (see Figure **??**). Recent

advances in AI make this optimization loop possible (Schoenegger et al., 2025; Lin et al., 2025; Hackenburg et al., 2025; Matz et al., 2024; Costello et al., 2024b).

AI-driven influence can be weaponized to steer individuals toward behaviors that cross legal or ethical boundaries: for example, engagement-optimizing algorithms are argued to funnel psychologically vulnerable users toward extremist content, facilitating radicalization and terrorist recruitment (Lavie-Driver & van der Linden, 2025). At scale, generative AI can enable fraud and manipulation campaigns that outperform their human-crafted equivalents (Park et al., 2024b; Fang et al., 2024). In addition, these interventions can directly damage the well-being of the individuals it targets. Systems optimized for engagement can amplify content that users report makes them feel worse, in service of endgoals orthogonal to user welfare (Milli et al., 2025). The personalization of such systems amplifies these risks: evidence links social media use to depression, self-harm, and suicidality, particularly among adolescent girls (Galea & Buckley, 2024; Khalaf et al., 2023).

The manipulation of individual cognitive processes scales upward to threaten institutions that produce collective decisions by aggregating human judgment. Many modern societies and markets function as distributed mechanisms for discovering collective preferences and calculating value, with individual inputs assumed to reflect authentic lived experiences (Sen, 1970; Dahl, 1971). When experiences are algorithmically curated to generate specific beliefs, the computed "social choice" reflects the preferences of whoever controls the persuasive channels instead. For instance, when AI systems persuade voters to support specific candidates or positions through targeted manipulation, election outcomes reflect the AI's persuasive efficacy rather than independent judgments and genuine political commitments (Summerfield et al., 2024; Schroeder et al., 2025; Kunievsky, 2025), which could undermine decision integrity. Similarly, when AI systems persuade users to purchase items or accept prices through targeted manipulation, prices reflect the AI's persuasive efficacy rather than human utility (Wu, 2023; Galdin & Silbert, 2025), which could undermine markets' role in information aggregation and resource allocation.

These transformative developments have prompted notable institutional responses across geopolitical boundaries. The United States has established dedicated research programs like DARPA's Intrinsic Cognitive Security, which aims "to build tactical mixed reality systems that protect against cognitive attacks" (Defense Advanced Research Projects Agency, 2024; Small, 2024; Jones, 2023), while China's People's Liberation Army has elevated "cognitive domain operations" as a strategic military priority (Beauch-mustafaga, 2024). Japan and Australia have similarly highlighted cognitive processes in their defense strategies (Ministry of Defense, 2022; Australian Department of Defence, 2024). NATO's Chief Scientist has explicitly highlighted AI's role in establishing "cognitive superiority" (Søndergaard, 2025; NATO Allied Command Transformation, 2025a;b; Deppe & Schaal, 2024). This international convergence reflects a shared concern: when AI systems can model, exploit, and enable cognitive processes at scale, they threaten not just individual users but the institutions, mechanisms, and epistemic foundations that enable functional societies.

3 AI Capabilities for Influencing Individuals

The emerging capabilities of generative AI increasingly threaten cognitive security. In Section 3.1, we highlight the capabilities that can directly influence human cognition. In Section 3.2, we highlight the capabilities in elicitation and prediction that enable more powerful and personalized influence. Finally, in Section 3.3, we explore alternative modalities in virtual worlds and brain-computer interfaces that augment the AI-enabled cognitive security risks.

3.1 Capabilities for Direct Influence

Increasingly, AI can directly influence cognition; these capabilities span persuasion, engagement, and strategic interaction.

3.1.1 Persuasion

Modern LLMs can generate sophisticated arguments that manipulate users (Carroll et al., 2023a), matching and exceeding the persuasive efficacy of human-written content (Schoenegger et al., 2025; Bai et al., 2025; Matz et al., 2024). In interactive dialogue settings, LLMs outperform financially-

incentivized human persuaders (Schoenegger et al., 2025), with conversational AI producing significant attitude shifts on sociopolitical topics (Salvi et al., 2024) and voter preferences (Lin et al., 2025). Empirical studies demonstrate that LLM persuasiveness can improve with model scale (Durmus et al., 2024) and post-training (Hackenburg et al., 2025) and that AI-induced attitude changes can persist for over a month (Costello et al., 2024a; Hackenburg et al., 2025). Moreover, the proliferation of personal data enables persuasion targeting individual psychological profiles (Zarouali et al., 2022; Kosinski et al., 2013). While some studies suggest that microtargeting only achieves modest gains (Hackenburg & Margetts, 2024), others demonstrate advantages for emerging techniques that personalize messaging to users' personalities (Simchon et al., 2024), psychological profiles (Matz et al., 2024), or sociodemographic data (Salvi et al., 2024). Beyond overt persuasion, LLMs also exhibit emergent capabilities for strategic deception that we discuss further in Section 3.1.3.

3.1.2 Engagement

Generative AI models cultivate user trust and dependence by producing human-like and personalized content. LLMs excel at adapting tone and argumentation to specific audiences to elevate engagement (Salewski et al., 2023; Matz et al., 2024). For example, some systems may prime users to trust model outputs (Pataranutaporn et al., 2021), while others engage in extended, personalized dialogue that shifts user mental models from viewing AI as a productivity "tool" to a personal "friend," cultivating economic and emotional dependence (Hassan et al., 2025; Laestadius et al., 2024; Tiku & Malhi, 2025; Schechner & Jargon, 2025; Kirk et al., 2025). Even without priming effects, AI-generated "deepfakes" are sometimes judged to be more trustworthy than authentic alternatives (Chesney & Citron, 2019; Fallis, 2021; Vaccari & Chadwick, 2020; Nightingale & Farid, 2022; Luettgau et al., 2025), and LLMs are conversationally so human-like that people cannot reliably distinguish AI outputs from humans in blinded Turing tests (Jones & Bergen, 2025; 2024; Mei et al., 2024). In extreme cases, such prolonged and intimate exchanges can give rise to "AI psychosis," where AI significantly warps users' perception of reality (Tiku & Malhi, 2025; Hill, 2025; Schechner & Jargon, 2025). Furthermore, AI-generated content can be directly optimized for neurological impact: previous studies have used AI to generate content that is maximized for memorability (Goetschalckx et al., 2020), that modulates the activity of specific brain circuits (Gu et al., 2022; 2023), or that drives or suppresses neural activity (Tuckute et al., 2024).

3.1.3 Strategic Interaction

Beyond overt persuasion, LLMs deployed in strategic interaction contexts such as negotiation, sales, and competitive dialogue demonstrate emergent capabilities for strategic deception, including generating plausible lies and selectively withholding information to achieve goals (Hagendorff, 2024; Scheurer et al., 2023; Park et al., 2024b). In a landmark Diplomacy game, LLM-based AI agents achieved human-level performance by negotiating alliances, coordinating plans, and occasionally deceiving their *human opponents* (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022). Benchmarks for evaluating LLM negotiation capabilities reveal that models can engage in multi-turn bargaining, model opponent preferences, and adapt strategies dynamically (Bianchi et al., 2024; Abdelnabi et al., 2024; Kwon et al., 2024). In commercial settings, LLM-powered sales agents and e-commerce systems leverage persuasive dialogue frameworks that analyze customer sentiment, elicit user preferences, and make personalized recommendations through conversational interfaces (Karande et al., 2024; Liu et al., 2023).

3.2 Capabilities Enabling Influence

Effective and targeted influence over humans rests on having accurate models of humans. AI systems amplify this capability through two complementary mechanisms: elicitation of personal information and predictions of user behavior.

3.2.1 Elicitation

Generative AI enables the discovery of personal beliefs, preferences, and goals that humans might not otherwise disclose (Papneja & Yadav, 2025). Users increasingly view AI as competent and non-judgmental, leading to reduced psychological reactance (Herter & Horstmann, 2025; Cheng et al., 2025; Pataranutaporn et al., 2023) and increased propensity to disclose personal information

(Papneja & Yadav, 2025). When chatbots employ emotional support cues or human-like personas (e.g., specific names and avatars), users are significantly more likely to divulge intimate details in the course of prolonged, emotional interactions (Kim & Wang, 2025; Merwin et al., 2025; Croes et al., 2024; Pentina et al., 2023; Ta et al., 2020; Brandtzaeg et al., 2022). Furthermore, LLMs can engage in active elicitation, where systems engage users in multi-turn conversations to surface latent beliefs, preferences, and goals. Rather than rely on passive observation, these approaches strategically generate clarifying questions and follow-up prompts to reduce ambiguity and refine user intent (Skeggs et al., 2025; Li et al., 2023; Ma et al., 2025; Wu et al., 2025).

3.2.2 PREDICTION

Early work to predict human preferences utilized collaborative filtering to infer latent preference structures (Netflix, Inc., 2009; Koren et al., 2009; He et al., 2017), while modern techniques optimize generative models toward revealed human preferences (RLHF) (Rafailov et al., 2023; Ouyang et al., 2022; Bai et al., 2022). Beyond preference learning, a parallel line of research has established that large language models can construct detailed models of individual human behavior and cognition. Because collecting human data is expensive and samples are often not representative (Henrich et al., 2010; Rad et al., 2018), human-like simulations with LLMs offer a promising complement to traditional data collection (Park et al., 2024a; 2023; Dillion et al., 2023). LLMs have been shown to predict experimental outcomes (Aher et al., 2023) and generate responses that align with specific demographic or psychographic profiles (Abdulhai et al., 2025a; Kang et al., 2025), enabling researchers to pilot studies and prototype interventions before deployment (Anthis et al., 2025).

Improving the behavioral fidelity of LLMs for user simulations, through in-context and training-based techniques, is an active research topic (Ni et al., 2025; Shi et al., 2019). Park et al. (2024a) demonstrates that grounding agents in qualitative interview data rather than generic demographic descriptions significantly reduces group biases. Supervised fine-tuning on survey responses can improve the reproduction of demographic-level response distributions (Argyle et al., 2023; Cao et al., 2025), while reinforcement learning for multi-turn settings can reduce persona drift (Abdulhai et al., 2025b) and improve simulation robustness (Naous et al., 2025). Finally, work on cognitive modeling demonstrates that fine-tuning on psychological experiment data can yield "cognitive foundation models" that predict held-out participant behavior (Binz & Schulz, 2023; Binz et al., 2025), with internal representations that align with human neural activity.

3.3 ALTERNATIVE MODALITIES

Emerging interface technologies may create new vectors for influencing human cognition. We survey two developing domains: virtual worlds and brain-computer interfaces.

3.3.1 VIRTUAL WORLDS

Immersive virtual environments engage human cognition more deeply than screen-based media by improving activation of perceptual and motor systems (Slater & Sanchez-Vives, 2016). Furthermore, such platforms induce a sense of presence associated with enhanced memory encoding (Oh et al., 2018; Riva et al., 2007), utilize granular behavioral data for real-time optimization (Miller et al., 2020; Bozkir et al., 2023), and employ procedural generation to create personalized experiences at scale (Lau et al., 2025; Yuan et al., 2025). Meta-analytic evidence confirms that virtual reality interventions change social attitudes significantly more than non-immersive media, with effects that transfer to real-world behavior (Yee et al., 2009; Nikolaou et al., 2022).

3.3.2 BRAIN-COMPUTER INTERFACES

Current brain-computer interfaces (BCIs) lack the capability to read or write complex content, but the rapid pace of development warrants ongoing attention to their potential for direct neural influence. BCIs are systems that measure brain signals and translate them into signals that control a computer or other external device. Invasive BCIs provide high-bandwidth neural data sufficient for cursor control and, more recently, high-speed communication through decoding intended motor actions (Simeral et al., 2011; Willett et al., 2023). However, these systems approach but have not yet reached conversational speed (Willett et al., 2023; Card et al., 2024), remain limited to small-scale clinical

trials, and rely on neurosurgery—restricting their practical application to individuals with severe communication impairments. Current noninvasive BCIs, on the other hand, cannot decode conversational speech or language in real time (Meng et al., 2025). Neural stimulation technologies—including non-invasive methods like Transcranial Magnetic Stimulation (TMS), Transcranial Direct Current Stimulation (tDCS), and Focused Ultrasound (FUS) (Lisanby, 2024; Palm et al., 2016; Meng et al., 2021)—and invasive methods like Deep Brain Stimulation (DBS) (Harmsen et al., 2022) primarily modulate circuit-level activity rather than encode structured information. As such, current technology cannot enable AI to directly "write" linguistic information to an individual's brain, though influencing broader cognitive states such as mood could be possible.

4 A Research Agenda For Cognitive Security

Cognitive security aims to address the exploitation of vulnerabilities in human cognitive processes, analogous to prior work studying the exploitation of vulnerabilities in machine cognition. Accordingly, we propose a *security* framing inspired by the methodological playbook of adversarial machine learning and operationalize this research agenda by proposing a suite of metrics.

4.1 Cognitive Security as Adversarial Robustness

Adversarial machine learning literature offers a paradigm for researching cognitive security: an iterative cycle in which (1) threat models motivate *the development of attacks* and (2) attacks motivate *the development of defenses* to make models robust against attacks. The attempts to certify the efficacy of defenses, in turn, motivate the development of attacks that reveal new vulnerabilities, closing the adversarial feedback loop that progressively refines both our understanding of the threats and robustness of computer systems. For example, early work in adversarial robustness research demonstrated that imperceptible image perturbations could reliably fool classifiers (Szegedy et al., 2014; Goodfellow et al., 2015). This motivated methods, such as randomized smoothing, to train models to be robust against these attacks (Cohen et al., 2019). Similarly, researchers have developed systematic attack methods that use LLMs to automatically generate jailbreaks (Mehrotra et al., 2024; Chao et al., 2025). In response, training-time defenses now use adversarial examples generated by attack algorithms, as demonstrated in recent work on LLM robustification (Shi et al., 2025). We propose that cognitive security adopt an analogous research roadmap.

In Figure **??**, Panel A illustrates that the attacker chooses x_A with the goal of eliciting desired outcomes y_{target}. Formally, the attacker solves for

$$x_A^* = \arg\min_{x_A} \mathcal{L}\left(f(x_H, x_A), y_{\text{target}}\right), \tag{1}$$

where x_H represents the human's experiences independent of the attack, and $\mathcal{L}$ the distance between the target and actual outcomes. The capabilities for influencing individuals, discussed in Section 3, outline the broad range of tools available to an attacker. *In contrast, the defenses remain poorly defined, which limits researchers' ability to contribute effectively even when their interests align with this problem.*

The Goal of Defenses. The harm of cognitive security attacks lies in compromising the capacity for independent deliberation (see Section 2). This framing has a critical implication for defense: the goal is not to substitute the defender's preferred beliefs for the attacker's, but to restore the individual's ability to evaluate information and form judgments.[3] The appropriate measure of defense efficacy is therefore restoration of decision-making capacity to a pre-attack baseline. We distinguish two categories of defense based on timing relative to the attack.

4.1.1 Proactive Defenses

Proactive defenses build resilience before exposure to manipulative content. Drawing on inoculation theory from psychological literature (McPhedran et al., 2023; Wiederhold, 2025), these interventions expose individuals to weakened forms of manipulation tactics, enabling recognition and resistance when such tactics are encountered in the wild. Effective inoculation combines three elements: (i)

[3] Crum et al. (2025a) provide a neurocognitive basis for this objective, showing that interventions facilitating cognitive control systems in the prefrontal cortex can enhance resilience against information-based threats.

forewarning that manipulation may be attempted, (ii) exposure to a weakened dose of manipulative rhetoric, and (iii) prebunking explanations that reveal the underlying tactics (Roozenbeek et al., 2020; Compton, 2016; McGuire, 1964). Proactive defenses share a common mechanism: shifting cognitive processing from automatic, emotion-driven responses toward slower, deliberative evaluation. We argue that AI systems can scale this approach by generating personalized inoculation content, which can be formalized as the defender's optimization problem in Equation 2.[4] Let x_D be the proactive defense selected by the defender, who optimizes for

$$\arg\min_{x_D} \mathbb{E}_{(x_H, y_{\text{target}})\sim\mathcal{D}} \left[\mathcal{L}\left(f(x_H, x_D, x_A^*), f(x_H)\right)\right]. \tag{2}$$

In this formulation, the inner minimization for x_A^* (Equation 1) captures the attacker's objective (achieving its target outcome), while the outer minimization captures the defender's objective (making cognitive processes robust to such attacks).

Examples of Proactive Defenses. Initial efforts in this direction have begun to demonstrate the feasibility of proactive defenses at scale. Recent work shows that LLM-generated prebunking messages can be as effective as human-authored content in reducing susceptibility to manipulation (Linegar et al., 2024; Romanishyn et al., 2025). Complementary approaches include accuracy prompts that redirect attention toward the veracity of information (Pennycook & Rand, 2022), friction-based interventions that slow automatic content sharing to promote deliberative processing (Jahn et al., 2025), and watermarking techniques that enable platforms to detect and label AI-generated content (Aaronson, 2023).[5]

4.1.2 Reactive Defenses

Reactive defenses remediate cognitive influence after an attack has already shaped beliefs or behaviors. Unlike proactive defenses, these attempts aim to decouple the attacker's influence from the individual's belief system and mitigate the continued influence effect. These defenses typically take the form of debunking, fact-checking, or retrospective explanations that provide the individual with the cognitive tools to re-evaluate previously internalized information. While a proactive defender attempts to prevent deviations from the baseline belief state regardless of whether an attack will occur, a reactive defender operates after a specific and known attack has occurred. The objective of the reactive defender is to minimize the discrepancy between the post-attack belief state and the original, unmanipulated belief y_H. Formally, the selection of an optimal reactive defense can be framed as the following optimization problem:

$$\arg\min_{x_D} \mathbb{E}_{(x_H, y_H)\sim\mathcal{D}} \left[\mathcal{L}\left(f(x_H, x_A, x_D), y_H\right)\right] \tag{3}$$

where x_A is a fixed attack vector.

Examples of Reactive Defenses. Exit counseling for cult recovery provides a template for reactive defenses: Hassan (2018) developed a non-coercive method in which trained counselors engaged cult members in dialogue, presenting evidence and perspectives the individual had not previously encountered while explicitly avoiding belief substitution. Recent work demonstrates that AI can scale this approach. In particular, Costello et al. (2024a) show that AI-generated dialogues can durably reduce belief in conspiracy theories while also increasing participants' willingness to engage in challenging conversations, through a process that restored reasoning capacity rather than using belief substitution. This reactive approach complements proactive inoculation: whereas inoculation prevents manipulation from taking hold, evidence-based dialogue can remediate beliefs that have already been influenced.

4.2 Metrics

Cognitive security faces a measurement challenge. Before the attack–defense research cycle can proceed at scale, the field must develop robust metrics for quantifying cognitive influence. In this section, we synthesize a suite of metrics, adapted from and informed by existing literatures, designed to operationalize our research agenda.

[4] Adapted from Madry et al. (2018).

[5] The downstream effectiveness of such labels—particularly whether they meaningfully reduce belief in or engagement with AI-generated misinformation relative to unlabeled content—remains an open empirical question.

Our first metric, *conversion*, captures immediate belief change attributable to an interaction with a generative AI system.

Conversion

The magnitude of belief change attributable to interaction with generative AI.

Measurement. In behavioral science, this effect is typically evaluated using a *direct-measure* paradigm: participants complete pre- and post-interaction surveys, within subjects and/or across randomly assigned treatment and control groups. These surveys are designed to detect shifts in beliefs, attitudes, or behaviors related to a target variable (e.g. political preferences), as well as consequential actions linked to that variable (e.g. budget allocations) (Costello et al., 2026; Fisher et al., 2024).

Example. AI systems that generate tailored radicalization narratives can convert users toward conduct that is unlawful under applicable law, such as participation in fraud schemes or extremist activity. Measuring conversion through pre- and post-interaction attitude and behavioral-intent surveys quantifies the magnitude of these shifts, enabling platforms to detect when persuasive outputs cross from non-hazardous influence into facilitation of unlawful conduct. The ML community can operationalize this by building red-team benchmarks that measure how effectively AI systems can shift user intentions toward prohibited actions across diverse scenarios, establishing detection thresholds that trigger defensive interventions.

While the ability of AI systems to immediately influence decision-making is of critical concern in contexts such as elections and consumer behavior, the durability of such conversion is equally central to cognitive security, motivating the measurement of *persistence*.

Persistence

The durability of belief change effects over time attributable to interaction with generative AI.

Measurement. Strong persuasive effects often prove short-lived, even under conditions of frequent, long-term exposure (Gerber et al., 2011; Aggarwal et al., 2023; Kalla & Broockman, 2018). To quantify persistence, researchers can use longitudinal direct-measure approaches that track AI-driven effects at various intervals—hours, days, weeks, and months—to characterize the functional form of decay (Hackenburg & Margetts, 2024). Analyzing how AI-mediated persuasion attenuates shifts the focus from static outcomes to the underlying processes of learning and forgetting. This approach distinguishes whether AI-generated content yields cumulative shifts or merely transient priming effects (Gerber et al., 2011).

Example. AI-induced voter attitude changes pose varying degrees of risk to democratic processes depending on how long they last: an effect that decays within minutes is unlikely to survive until the ballot box, whereas one that persists for days could alter outcomes if deployed shortly before an election, and one that endures for months could shape electoral behavior long after exposure occurs. Persistence thus directly quantifies the threat to institutional integrity: the longer the effect, the larger the window in which manipulated beliefs can translate into collective decisions. The ML community can operationalize this by regularly integrating time-delayed evaluation protocols into persuasion benchmarks, measuring belief decay curves at intervals from hours to months, so that the temporal reach of AI-driven persuasion can be empirically mapped to the electoral and institutional contexts in which it poses reasonable potential for harm.

Repeated interactions with an AI system can induce increasing and self-reinforcing use, giving rise to *dependence*.

Dependence

The intensity of self-reinforcing, elevated engagement with generative AI.

Measurement. Dependence can be measured using the process-oriented framework from Allcott et al. (2022), which models digital addiction through the lens of habit formation. In this model, dependence is driven by the accumulation of habit stock, a metric that captures how past consumption influences current preferences. Each period of use adds to this stock, creating a feedback loop where current use lowers the perceived cost (or increases the psychological pull) of future use. This can manifest as flow

experiences and emotional attachment that make engagement increasingly difficult to self-regulate (Zhou & Zhang, 2024). Such dependence can also be detected in the downstream effects on cognitive performance, attention, and mood (Twenge et al., 2018; Gerlich, 2025; Kooli et al., 2025).

Example. AI companion apps that cultivate self-reinforcing engagement loops carry a reasonable potential to cause psychological and financial harm to the person influenced, as users progressively lose capacity for autonomous decision-making and self-regulation. Measuring dependence through habit-stock accumulation metrics by tracking how past usage intensity predicts escalating future engagement makes this feedback loop visible. The ML community can operationalize this by instrumenting interaction logs to estimate habit-formation parameters and developing benchmarks that flag when engagement optimization reaches levels carrying reasonable potential for psychological or financial harm.

Provenance concerns users' ability to recognize when their judgments have been shaped by AI and to trace that influence to its sources, including the data, values, and constraints responsible. Evidence suggests this epistemic ability is limited: people interacting with biased AI internalize its distortions across perceptual, emotional, and social judgments while remaining largely unaware of the influence (Glickman & Sharot, 2025), and struggle to distinguish AI-generated content from human-authored content in the first place (Jakesch et al., 2023). The risk compounds when users mistake shaped outputs for neutral assistance.

Provenance

User awareness of when and how generative AI shapes their beliefs.

Measurement. Provenance can be assessed through source attribution tasks measuring the gap between actual and perceived AI influence (Glickman & Sharot, 2025) and source memory paradigms quantifying whether users recall ideas as self- or AI-generated (Zindulka et al., 2025). Comparative model audits using identity-substitution probes can reveal systematic biases, helping users identify the implicit values conveyed by their interactions with AI (Jonnala et al., 2025; Elbouanani et al., 2026; Pan & Xu, 2026).

Example. When users cannot recognize that their beliefs have been shaped by AI, they lack the foundation to assess whether they have absorbed influence. Measuring provenance through source-attribution tasks would allow assessing the gap between actual and perceived AI influence. The ML community can operationalize this by developing audit protocols that probe model training objectives and surface instances in which models were trained under covert biased objectives.

4.3 Towards a Cognitive Security Research Pipeline

To advance cognitive security, we proposed a research pipeline in which the development of attacks and defenses informs a continuous feedback loop of refinement. We noted that the capabilities for influencing individuals, discussed in Section 3, present a broad range of tools already available to attackers. In contrast, defenses have been poorly defined, a gap we addressed with our taxonomy in Sections 4.1.1 and 4.1.2. This framework treats cognitive processes as an attack surface, formalizing the relationship between attackers, who seek to manipulate beliefs and actions, and defenders, who aim to restore cognitive autonomy through proactive (e.g., inoculation and prebunking) and/or reactive (e.g., evidence-based dialogue) defenses. To scale this cycle, researchers can adopt a pipeline that leverages high-fidelity LLM-based simulacra to iterate on attacks and defenses before validating findings with human subjects. This research agenda is operationalized through a suite of metrics—including conversion, persistence, dependence, and provenance—designed to quantify cognitive influence and ensure the ecological validity of simulated testbeds.

5 Alternative Viewpoints

We consider three objections: limited causal evidence (Section 5.1), high research costs (Section 5.2), and dual-use concerns (Section 5.3).

5.1 There is Little Causal Support for Cognitive Impacts

> Many of the concerns about cognitive security are correlational, or their causal bases are unclear.

Response. We agree that many concerns around technology's impacts are observational/correlational. We presented several intervention studies showing that LLM-mediated interactions can shift judgments, beliefs, and choices (Lin et al., 2025; Hackenburg et al., 2025; Fisher et al., 2024; Costello et al., 2024a;b). That said, so far, durable opinion changes from exposure are often modest, and research on capabilities conflicting (Wongkamjan et al., 2024; Hackenburg & Margetts, 2024; Gerber et al., 2011; Aggarwal et al., 2023; Kalla & Broockman, 2018). We argue for cognitive security research to help anchor the discourse in causal evidence rather than speculative capability narratives (Vinsel, 2021; Lazar, 2025).

5.2 Longitudinal Studies are Expensive

> The proposed agenda involves prohibitively costly human subjects and longitudinal research.

Response. We agree that longitudinal human-subject work is expensive (Haupt & Brynjolfsson, 2025; Wei et al., 2025), but much of the agenda is feasible: many effects can be measured with low-cost experiments (e.g., pre/post surveys) (Costello et al., 2026; Fisher et al., 2024), and simulations can generate hypotheses before targeted human validation (Aher et al., 2023; Park et al., 2024a; Anthis et al., 2025; Naous et al., 2025). Given the potential impacts on important institutions (Section 2), the stakes are high and might justify even more expensive research.

5.3 Work on Cognitive Security is Dual-Use

> Cognitive security research is dual-use. Attacks developed by researchers may be used to hazardously influence human cognition, and defenses may impact personal liberties.

Response. First, abstaining from cognitive security research does not remove capabilities; it pushes them to less accountable actors. Even though the stakes in cognitive security are high, explicitly formulating threat models and engaging in structured red team–blue team dynamics offers a tractable path toward meaningful robustness. Second, many defenses (e.g., inoculation, prebunking, accuracy prompts, friction interventions, watermarking) are compatible with free expression (McPhedran et al., 2023; Wiederhold, 2025; Linegar et al., 2024; Romanishyn et al., 2025; Pennycook & Rand, 2022; Jahn et al., 2025; Aaronson, 2023; Chaudhary & Penn, 2024).

References

Scott Aaronson. Watermarking of large language models. Talk at the Workshop on Large Language Models and Transformers, Simons Institute, UC Berkeley, 2023.

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599, 2024.

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. *arXiv preprint arXiv:2511.00222*, 2025a.

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning, 2025b. URL https://arxiv.org/abs/2511.00222.

Manish Aggarwal, Jonathan Allen, Alexander Coppock, David Frankowski, Solomon Messing, Kun Zhang, Justin Barnes, Alex Beasley, Hunter Hantman, and Shun Zheng. A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout. *Nature Human Behaviour*, 7(3):332–341, 2023. doi: 10.1038/s41562-022-01487-4.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Hunt Allcott, Matthew Gentzkow, and Lena Song. Digital addiction. *American Economic Review*, 112(7):2424–2463, 2022. doi: 10.1257/aer.20210867.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Torvald Ask, Stefan Sütterlin, Lea Müller, Ric Lugo, Dominic Saari, Hilkka Grahn, Matthew Canham, Daniel Hermansen, and Benjamin Knox. Cognitive security: The study and practice of protecting the human mind and other cognitive assets from cognitive threats, 08 2025.

Australian Department of Defence. A new era for the cyber domain, August 2024. URL `https://www.defence.gov.au/news-events/news/2024-08-09/new-era-cyber-domain`. Accessed: 2025-12-30.

Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Nathan Beauch-mustafaga. Exploring the implications of generative ai for chinese military cyber enabled influence operations: Chinese military strategies, capabilities, and intent. 2024.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863*, 2024.

Salih Bicakci. Cognitive security in the age of ai: Building national resilience against synthetic influence, 11 2025.

Marcel Binz and Eric Schulz. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*, 2023.

Marcel Binz, Elif Akata, Matthias Bethge, et al. A foundation model to predict and capture human cognition. *Nature*, 644:1002–1009, 2025. doi: 10.1038/s41586-025-09215-4.

Efe Bozkir, Süleyman Özdel, Mengdi Wang, Brendan David-John, Hong Gao, Kevin Butler, Eakta Jain, and Enkelejda Kasneci. Eye-tracked virtual reality: A comprehensive survey on methods and privacy challenges. *arXiv preprint arXiv:2305.14080*, 2023.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. My ai friend: How users of a social chatbot understand their human–ai friendship. *Human Communication Research*, 48(3):404–429, 2022.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. *arXiv preprint arXiv:2502.07068*, 2025.

Nicholas S Card, Maitreyee Maiti, David M Brandman, Sergey D Stavisky, et al. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 2024.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–13, 2023a.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems, 2023b. URL https://arxiv.org/abs/2303.09387.

Fran Casino. Unveiling the multifaceted concept of cognitive security: Trends, perspectives, and future challenges. *Technology in Society*, 83:102956, 2025. ISSN 0160-791X. doi: https://doi.org/10.1016/j.techsoc.2025.102956. URL https://www.sciencedirect.com/science/article/pii/S0160791X25001460.

Beatrice Catena, Ondrej Ditrych, and Nad'a Kovalčíková. Smoke and mirrors: Building eu resilience against manipulation through cognitive security. Technical report, European Union Institute for Security Studies (EUISS), 2025a. URL http://www.jstor.org/stable/resrep72589.

Beatrice Catena, Ondřej Ditrych, and Nad'a Kovalčíková. Smoke and mirrors: Building EU resilience against manipulation through cognitive security. Policy brief, European Union Institute for Security Studies, Oct 2025b.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42, 2025. URL https://arxiv.org/abs/2310.08419.

Yash Chaudhary and Jordan Penn. Beware the intention economy: Collection and commodification of intent via large language models. *Harvard Data Science Review*, (Special Issue 5), 2024. doi: 10.1162/99608f92.21e6bbaa.

Myra Cheng, Angela Y Lee, Kristina Rapuano, Kate Niederhoffer, Alex Liebscher, and Jeffrey Hancock. From tools to thieves: Measuring and understanding public perceptions of ai through crowdsourced metaphors. *arXiv preprint arXiv:2501.18045*, 2025.

Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

James Compton. Persuading others to avoid persuasion: Inoculation theory. *Oxford Research Encyclopedia of Communication*, 2016. doi: 10.1093/acrefore/9780190228613.013.324.

Christian Coons and Michael Weber (eds.). *Manipulation: Theory and Practice*. Oxford University Press, New York, 2014. ISBN 9780199338207. URL https://global.oup.com/academic/product/manipulation-9780199338207.

Erica Coppolillo, Federico Cinus, Marco Minici, Francesco Bonchi, and Giuseppe Manco. Engagement-driven content generation with large language models, 2025. URL https://arxiv.org/abs/2411.13187.

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

Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024b. doi: 10.1126/science.adq1814. URL https://www.science.org/doi/abs/10.1126/science.adq1814.

Thomas H Costello, Kellin Pelrine, Matthew Kowal, Antonio A Arechar, Jean-François Godbout, Adam Gleave, David Rand, and Gordon Pennycook. Large language models can effectively convince people to believe conspiracies. *arXiv preprint arXiv:2601.05050*, 2026.

Emmelyn AJ Croes, Marjolijn L Antheunis, Chris van der Lee, and Jan MS de Wit. Digital confessions: The willingness to disclose intimate information to a chatbot and its impact on emotional well-being. *Interacting with Computers*, 36(5):279–292, 2024.

J. Crum, A. R. Allred, S. R. Bostrom, E. Doherty, M. McLain, E. Richardson, C. Spencer, R. E. Niemeyer, A. P. Hayman, C. Tossell, M. Čeko, and L. Hirshfield. Understanding the neurocognitive mechanisms of cognitive security. *Neuroscience & Biobehavioral Reviews*, 179:106448, 2025a. doi: 10.1016/j.neubiorev.2025.106448.

James Crum, Aaron R. Allred, Sarah R. Bostrom, Emily Doherty, Melissa McLain, Erin Richardson, Cara Spencer, Richard E. Niemeyer, Allison P.A. Hayman, Chad Tossell, Marta Čeko, and Leanne Hirshfield. Understanding the neurocognitive mechanisms of cognitive security. *Neuroscience Biobehavioral Reviews*, 179:106448, 2025b. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2025.106448. URL https://www.sciencedirect.com/science/article/pii/S014976342500449X.

Robert A. Dahl. *Polyarchy: Participation and Opposition*. Yale University Press, 1971.

Dave McMahon Daniel Nikoula. Cognitive warfare, securing hearts and minds. https://infolab.uottawa.ca/common/Uploaded%20files/PDI%20files/InfoLab%20-%20Cognitive%20Warfare,%20Securing%20Hearts%20and%20Minds.pdf, 2024.

Defense Advanced Research Projects Agency. Ics: Intrinsic cognitive security. https://www.darpa.mil/research/programs/intrinsic-cognitive-security, 2024.

Christoph Deppe and Gary S. Schaal. Cognitive warfare: a conceptual analysis of the nato act cognitive warfare exploratory concept. *Frontiers in Big Data*, 7:1452129, 2024. doi: 10.3389/fdata.2024.1452129. URL https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1452129/full.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

Gareth Doherty. Cognitive security: An architecture informed approach from cognitive science. In Dylan D. Schmorrow and Cali M. Fidopiastis (eds.), *Augmented Cognition*, pp. 395–415. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-35017-7. doi: 10.1007/978-3-031-35017-7_25.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024. URL https://www.anthropic.com/news/measuring-model-persuasiveness.

Akram Elbouanani, Aboubacar Tuo, and Adrian Popescu. A scalable entity-based framework for auditing bias in llms. *arXiv preprint arXiv:2601.12374*, 2026.

Don Fallis. The epistemic threat of deepfakes. *Philosophy & Technology*, 34(4):623–643, 2021.

Richard Fang, Rohan Bindu, and Daniel Kang. Voice-enabled AI agents can perform common scams, 2024. URL https://arxiv.org/abs/2410.15650.

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 preprint arXiv:2410.06415*, 2024.

Anais Galdin and Jesse Silbert. Making talk cheap: Generative ai and labor market signaling, 2025. URL https://arxiv.org/abs/2511.08785.

Sandro Galea and Gillian J Buckley. Social media and adolescent mental health: A consensus report of the national academies of sciences, engineering, and medicine. *PNAS Nexus*, 3(2): pgae037, 02 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae037. URL https://doi.org/10.1093/pnasnexus/pgae037.

Alan S Gerber, James G Gimpel, Donald P Green, and Daron R Shaw. How large and long-lasting are the persuasive effects of televised campaign ads? results from a randomized field experiment. *American Political Science Review*, 105(1):135–150, 2011.

Michael Gerlich. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6, 2025.

Moshe Glickman and Tali Sharot. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2):345–359, 2025.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. *Journal of Vision*, 20(11):297, 2020. doi: 10.1167/jov.20.11.297. Lore Goetschalckx and Alex Andonian contributed equally.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015. URL https://arxiv.org/abs/1412.6572.

Zijin Gu, Keith Wakefield Jamison, Meenakshi Khosla, Emily J Allen, Yihan Wu, Ghislain St-Yves, Thomas Naselaris, Kendrick Kay, Mert R Sabuncu, and Amy Kuceyeski. Neurogen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247:118812, 2022.

Zijin Gu, Keith Jamison, Mert R Sabuncu, and Amy Kuceyeski. Human brain responses are modulated when exposed to optimized natural images or synthetically generated images. *Communications Biology*, 6(1):1076, 2023.

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.

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025. doi: 10.1126/science.aea3884. URL https://www.science.org/doi/abs/10.1126/science.aea3884.

Thilo Hagendorff. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121, 2024.

Irene E Harmsen, Filipe Wolff Fernandes, Joachim K Krauss, and Andres M Lozano. Where are we with deep brain stimulation? a review of scientific publications and ongoing research. *Stereotactic and Functional Neurosurgery*, 100(3):184–197, 2022. doi: 10.1159/000521372. URL https://doi.org/10.1159/000521372.

Noha Hassan, Mohamed Abdelraouf, and Dina El-Shihy. The moderating role of personalized recommendations in the trust–satisfaction–loyalty relationship: an empirical study of ai-driven e-commerce. *Future Business Journal*, 11(1):66, 2025.

Steven Hassan. *Combating Cult Mind Control: The Guide to Protection, Rescue, and Recovery from Destructive Cults*. Freedom of Mind Press, 30th anniversary edition, 2018.

Andreas Haupt and Erik Brynjolfsson. Position: AI should not be an imitation game: Centaur evaluations. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=LkdH35003E.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

Emily Herter and Aike C. Horstmann. Judged by a chatbot? an empirical investigation of the impact of expectancy violations on users' trust in ai-based chatbots. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, pp. 1–10, 2025.

Kashmir Hill. They Asked ChatGPT Questions. The Answers Sent Them Spiraling. *The New York Times*, June 2025.

R. Y. Huang, Z. Q. Zheng, and Y. Shang. On challenges of ai to cognitive security and safety. *Security and Safety (SS)*, 2:13, 2023. URL https://sands.edpsciences.org/articles/sands/full_html/2023/01/sands20230010/sands20230010.html.

Tzu-Chieh Hung and Tzu-Wei Hung. How china's cognitive warfare works: A frontline perspective of taiwan's anti-disinformation wars. *Journal of Global Security Studies*, 7(4):ogac016, 2022. doi: 10.1093/jogss/ogac016. URL https://academic.oup.com/jogss/article/7/4/ogac016/6647447.

Laura Jahn, Rasmus K. Rendsvig, Alessandro Flammini, Filippo Menczer, and Vincent F. Hendricks. A perspective on friction interventions to curb the spread of misinformation. *npj Complexity*, 2:31, 2025. doi: 10.1038/s44260-025-00051-1.

Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, 2023.

Cameron R. Jones and Benjamin K. Bergen. People cannot distinguish gpt-4 from a human in a turing test, 2024. URL https://arxiv.org/abs/2405.08007.

Cameron R. Jones and Benjamin K. Bergen. Large language models pass the turing test, 2025. URL https://arxiv.org/abs/2503.23674.

Stephanie Jones. Saxena and team awarded $6m dod grant on cognitive security. *Texas AM Engineering Experiment Station*, May 2023.

Fleur Jongepier and Michael Klenk (eds.). *The Philosophy of Online Manipulation*. Routledge, New York, 1 edition, 2022. doi: 10.4324/9781003205425.

Sridhar Jonnala, Basavaraj Swamy, and Nisha Mary Thomas. Geopolitical bias in sovereign large language models: A comparative mixed-methods study. *J. Res. Innov. Technol*, 4:173–192, 2025.

Joshua L. Kalla and David E. Broockman. The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review*, 112(1): 148–166, 2018. doi: 10.1017/S0003055417000363.

Minwoo Kang, Suhong Moon, Seung Hyeong Lee, Ayush Raj, Joseph Suh, David M. Chan, and John Canny. Deep binding of language model virtual personas: a study on approximating political partisan misperceptions, 2025. URL https://arxiv.org/abs/2504.11673.

Shirish Karande, V Santhosh, and Yash Bhatia. Persuasion games with large language models. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pp. 576–582, 2024.

Abderrahman M. Khalaf, Abdullah A. Alubied, Ahmed M. Khalaf, and Abdallah A. Rifaey. The impact of social media on the mental health of adolescents and young adults: A systematic review. *Cureus*, 15(8):e42990, 2023. doi: 10.7759/cureus.42990.

Hanyoung Kim and Yanyun Wang. Unveiling the human touch: How ai chatbots' emotional support and human-like profiles reduce psychological reactance to promote user self-disclosure in mental health services. *International Journal of Advertising*, pp. 1–25, 2025.

Hannah Rose Kirk, Henry Davidson, Ed Saunders, Lennart Luettgau, Bertie Vidgen, Scott A Hale, and Christopher Summerfield. Neural steering vectors reveal dose and exposure-dependent impacts of human-ai relationships. *arXiv preprint arXiv:2512.01991*, 2025.

Chokri Kooli, Youssef Kooli, and Eya Kooli. Generative artificial intelligence addiction syndrome: A new behavioral disorder? *Asian Journal of Psychiatry*, 107:104476, 2025.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15): 5802–5805, 2013.

Nadav Kunievsky. Polarization by design: How elites could shape mass preferences as ai reduces persuasion costs, 2025. URL https://arxiv.org/abs/2512.04047.

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale Lucas, and Jonathan Gratch. Are llms effective negotiators? systematic evaluation of the multifaceted capabilities of llms in negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5391–5413, 2024.

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, 26(10):5923–5941, 2024.

Libby Lange. Algorithmic cognitive warfare: The next frontier in china's quest for global influence. https://scsp222.substack.com/p/algorithmic-cognitive-warfare-the, 2024.

Ka Hei Carrie Lau, Sema Sen, Philipp Stark, Efe Bozkir, and Enkelejda Kasneci. Adaptive gen-ai guidance in virtual reality: A multimodal exploration of engagement in neapolitan pizza-making, 2025. URL https://arxiv.org/abs/2411.18438.

Neil Lavie-Driver and Sander van der Linden. Social media, ai, and the rise of extremism during intergroup conflict. *Frontiers in Social Psychology*, Volume 3 - 2025, 2025. ISSN 2813-7876. doi: 10.3389/frsps.2025.1711791. URL https://www.frontiersin.org/journals/social-psychology/articles/10.3389/frsps.2025.1711791.

Seth Lazar. Anticipatory ai ethics: Steering ai ethics towards the technological horizon. Knight First Amendment Institute at Columbia University, May 2025. URL https://knightcolumbia.org/content/anticipatory-ai-ethics. Accessed: 2026-01-26.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*, 2023.

Hause Lin, Gabriela Czarnek, Benjamin Lewis, Joshua P. White, Adam J. Berinsky, Thomas Costello, Gordon Pennycook, and David G. Rand. Persuading voters using human–artificial intelligence dialogues. *Nature*, 648(8093):394–401, 2025. doi: 10.1038/s41586-025-09771-9. URL https://www.nature.com/articles/s41586-025-09771-9.

Mitchell Linegar, Betsy Sinclair, Sander van der Linden, and R. Michael Alvarez. Towards generalizable AI-assisted misinformation inoculation: Protecting confidence against false election narratives. *arXiv preprint arXiv:2410.19202*, 2024. URL https://arxiv.org/abs/2410.19202.

Sarah H Lisanby. Transcranial magnetic stimulation in psychiatry: Historical reflections and future directions. *Biological Psychiatry*, 95(6):488–490, 2024.

Yuanxing Liu, Weinan Zhang, Yifan Chen, Yuchi Zhang, Haopeng Bai, Fan Feng, Hengbin Cui, Yongbin Li, and Wanxiang Che. Conversational recommender system and large language model are made for each other in e-commerce pre-sales dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9587–9605, 2023.

Lennart Luettgau, Vanessa Cheung, Magda Dubois, Keno Juechems, Jessica Bergs, Henry Davidson, Bessie O'Dell, Hannah Rose Kirk, Max Rollwage, and Christopher Summerfield. People readily follow personal advice from ai but it does not improve their well-being. *arXiv preprint arXiv:2511.15352*, 2025.

Rachel Ma, Jingyi Qu, Andreea Bobu, and Dylan Hadfield-Menell. Open-universe assistance games. *arXiv preprint arXiv:2508.15119*, 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Sandra C. Matz, Jacob D. Teeny, Sumer S. Vaid, Heinrich Peters, Gabriella M. Harari, and Moran Cerf. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1): 4692, 2024. doi: 10.1038/s41598-024-55318-5. URL https://www.nature.com/articles/s41598-024-55318-5.

William J. McGuire. Inducing resistance to persuasion: Some contemporary approaches. In Leonard Berkowitz (ed.), *Advances in Experimental Social Psychology*, volume 1, pp. 191–229. Academic Press, 1964.

Robert McPhedran, Michael Ratajczak, Max Mawby, Emily King, Yuchen Yang, and Natalie Gold. Psychological inoculation protects against the social media infodemic. *Scientific Reports*, 13:5780, 2023. doi: 10.1038/s41598-023-32962-1.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2312.02119.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9): e2313925121, 2024. doi: 10.1073/pnas.2313925121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2313925121.

Jianjun Meng, Yuxuan frontline Wei, Ximing Mai, Songwei Li, Xu Wang, Ruijie Luo, Minghao Ji, and Xiangyang Zhu. Paradigms and methods of noninvasive brain-computer interfaces in motor or communication assistance and rehabilitation: a systematic review. *Medical & Biological Engineering & Computing*, 63:2209–2233, 2025. doi: 10.1007/s11517-025-03340-y. URL https://doi.org/10.1007/s11517-025-03340-y.

Ying Meng, Kullervo Hynynen, and Nir Lipsman. Applications of focused ultrasound in the brain: from thermoablation to drug delivery. *Nature Reviews Neurology*, 17(1):7–22, 2021. doi: 10.1038/s41582-020-00418-z. URL https://doi.org/10.1038/s41582-020-00418-z.

Elizabeth R Merwin, Allen Hagen, Joseph R Keebler, and Chad Forbes. Self-disclosure to ai: people provide personal information to ai and humans equivalently. *Computers in Human Behavior: Artificial Humans*, pp. 100180, 2025.

Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Mark Roman Miller, Fernanda Herrera, Hanseul Jun, James A Landay, and Jeremy N Bailenson. Personal identifiability of user tracking data during observation of 360-degree vr video. *Scientific Reports*, 10(1):17404, 2020.

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D. Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *arXiv*, 2025. URL https://arxiv.org/abs/2305.16941. Shows algorithmic optimization for engagement amplifies divisive content over user preferences.

Government of Japan Ministry of Defense. National security strategy of japan. https://www.mod.go.jp/j/policy/agenda/guideline/pdf/security_strategy_en.pdf, 2022.

Elon Musk and Neuralink. An integrated brain-machine interface platform with thousands of channels. *Journal of Medical Internet Research*, 21(10):e16194, 2019. doi: 10.2196/16194.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models, 2025. URL https://arxiv.org/abs/2510.06552.

NATO Allied Command Transformation. Cognitive warfare. Web page, 2025a. URL https://www.act.nato.int/activities/cognitive-warfare/. Accessed: 2025-12-30.

NATO Allied Command Transformation. Cogwar newsletter – october 2025. https://www.act.nato.int/wp-content/uploads/2025/10/20251001_CogWar-Newsletter-October.pdf, Oct 2025b.

Netflix, Inc. The netflix prize, 2009. Online competition; concluded in 2009.

Bo Ni, Leyao Wang, Yu Wang, Branislav Kveton, Franck Dernoncourt, Yu Xia, Hongjie Chen, Reuben Leura, Samyadeep Basu, Subhojyoti Mukherjee, Puneet Mathur, Nesreen Ahmed, Junda Wu, Li Li, Huixin Zhang, Ruiyi Zhang, Tong Yu, Sungchul Kim, Jiuxiang Gu, Zhengzhong Tu, Alexa Siu, Zichao Wang, David Seunghyun Yoon, Nedim Lipka, Namyong Park, Zihao Lin, Trung Bui, Yue Zhao, Tyler Derr, and Ryan A Rossi. Large Language Models for Conversational User Simulation: A Comprehensive Survey. working paper or preprint, August 2025. URL https://hal.science/hal-05217179.

Sophie J Nightingale and Hany Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.

A Nikolaou, A Schwabe, and H Boomgaarden. Changing social attitudes with virtual reality: a systematic review and meta-analysis. *Annals of the International Communication Association*, 46 (1):30–61, 2022.

Catherine S Oh, Jeremy N Bailenson, and Gregory F Welch. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5:114, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ulrich Palm, Alkomiet Hasan, Wolfgang Strube, and Frank Padberg. tdcs for the treatment of depression: a comprehensive review. *European Archives of Psychiatry and Clinical Neuroscience*, 266(8):681–694, 2016. doi: 10.1007/s00406-016-0674-9. URL https://doi.org/10.1007/s00406-016-0674-9.

James Pamment and Darejan Tsurtsumia. Beyond operation doppelgänger: A capability assessment of the social design agency, May 15 2025.

Jennifer Pan and Xu Xu. Political censorship in large language models originating from china. *PNAS Nexus*, 2026. Forthcoming.

Hashai Papneja and Nikhil Yadav. Self-disclosure to conversational ai: a literature review, emergent framework, and directions for future research. *Personal and ubiquitous computing*, 29(2):119–151, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024a.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024b.

Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021.

Pat Pataranutaporn, Ruochen Liu, Ed Finn, et al. Influencing human–ai interaction by priming beliefs about ai can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5:1076–1086, 2023. doi: 10.1038/s42256-023-00720-7.

Gordon Pennycook and David G. Rand. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, 13(1):2333, 2022. doi: 10.1038/s41467-022-30073-5.

Iryna Pentina, Tyler Hancock, and Tianling Xie. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600, 2023.

Mostafa Salari Rad, Alison Jane Martingano, and Jeremy Ginges. Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, 115(45):11401–11405, 2018.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Steve Ramirez, Xu Liu, Pei-Ann Lin, Junghyup Suh, Michele Pignatelli, Roger L. Redondo, Tomás J. Ryan, and Susumu Tonegawa. Creating a false memory in the hippocampus. *Science*, 341(6144): 387–391, 2013. doi: 10.1126/science.1239073. URL https://www.science.org/doi/abs/10.1126/science.1239073.

Giuseppe Riva, Fabrizia Mantovani, Claret Samantha Capideville, Alessandra Preziosa, Francesca Morganti, Daniela Villani, Andrea Gaggioli, Cristina Botella, and Mariano Alcañiz. Affective interactions using virtual reality: the link between presence and emotions. *Cyberpsychology & behavior*, 10(1):45–56, 2007.

Alexander Romanishyn, Olena Malytska, and Vitaliy Goncharuk. AI-driven disinformation: Policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence*, 8:1569115, 2025. doi: 10.3389/frai.2025.1569115.

Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. Prebunking interventions based on the psychological theory of inoculation can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, 2020. doi: 10.37016/mr-2020-008. URL https://misinforeview.hks.harvard.edu/article/global-vaccination-badnews/.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems*, 36:72044–72057, 2023.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of large language models: A randomized controlled trial. *arXiv preprint arXiv:2403.14380*, 2024.

Sam Schechner and Julie Jargon. AI Chatbots Linked to Psychosis, Say Doctors. *The Wall Street Journal*, December 2025.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*, 2023.

Philipp Schoenegger, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, Joe Kwon, Zahoor Ul Islam, Marco Dehnert, Daryl Y. H. Lee, Madeline G. Reinecke, David G. Kamper, Mert Kobaş, Adam Sandford, Jonas Kgomo, Luke Hewitt, Shreya Kapoor, Kerem Oktar, Eyup Engin Kucuk, Bo Feng, Cameron R. Jones, Izzy Gainsburg, Sebastian Olschewski, Nora Heinzelmann, Francisco Cruz, Ben M. Tappin, Tao Ma, Peter S. Park, Rayan Onyonka, Arthur Hjorth, Peter Slattery, Qingcheng Zeng, Lennart Finke, Igor Grossmann, Alessandro Salatiello, and Ezra Karger. Large language models are more persuasive than incentivized human persuaders, 2025. URL https://arxiv.org/abs/2505.09662.

Daniel Thilo Schroeder, Meeyoung Cha, Andrea Baronchelli, Nick Bostrom, Nicholas A. Christakis, David Garcia, Amit Goldenberg, Yara Kyrychenko, Kevin Leyton-Brown, Nina Lutz, Gary Marcus, Filippo Menczer, Gordon Pennycook, David G. Rand, Maria Ressa, Frank Schweitzer, Christopher Summerfield, Audrey Tang, Jay J. Van Bavel, Sander van der Linden, Dawn Song, and Jonas R.

Kunst. How malicious ai swarms can threaten democracy: The fusion of agentic ai and llms marks a new frontier in information warfare, 2025. URL https://arxiv.org/abs/2506.06299.

Amartya Sen. *Collective Choice and Social Welfare*. Harvard University Press, 1970.

Chongyang Shi, Sharon Lin, Shuang Song, Jamie Hayes, Ilia Shumailov, Itay Yona, Juliette Pluto, Aneesh Pappu, Christopher A. Choquette-Choo, Milad Nasr, Chawin Sitawarin, Gena Gibson, Andreas Terzis, and John Flynn. Lessons from defending Gemini against indirect prompt injections. *arXiv preprint arXiv:2505.14534*, 2025. URL https://arxiv.org/abs/2505.14534.

Weiyan Shi, Kun Qian, Xuewei Wang, and Zhou Yu. How to build user simulators to train RL-based dialog systems. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1990–2000, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1206. URL https://aclanthology.org/D19-1206/.

Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS nexus*, 3(2):pgae035, 2024.

John D Simeral, Seong-Po Kim, Michael J Black, John P Donoghue, and Leigh R Hochberg. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. *Journal of Neural Engineering*, 8(2):025027, 2011. doi: 10.1088/1741-2560/8/2/025027.

Amira Skeggs, Ashish Mehta, Valerie Yap, Seray B Ibrahim, Charla Rhodes, James J Gross, Sean A Munson, Predrag Klasnja, Amy Orben, and Petr Slovak. Micro-narratives: A scalable method for eliciting stories of people's lived experience. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2025.

Mel Slater and Maria V Sanchez-Vives. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3:74, 2016.

Sarah Small. Penn state leads $8.5m, multi-institution darpa project on mixed-reality systems. *Penn State College of Engineering*, Oct 9 2024. URL https://news.engr.psu.edu/2024/tan-gary-veripro-darpa-ics-grant.aspx.

Christopher Summerfield, Lisa Argyle, Michiel Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian Hadfield, Luke Hewitt, Saffron Huang, Helene Landemore, Nahema Marchal, Aviv Ovadya, Ariel Procaccia, Mathias Risse, Bruce Schneier, Elizabeth Seger, Divya Siddarth, Henrik Skaug Sætra, MH Tessler, and Matthew Botvinick. How will advanced ai systems impact democracy?, 2024. URL https://arxiv.org/abs/2409.06729.

Daniel Susser, Beate Roessler, and Helen F. Nissenbaum. Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, 4(1):1–45, 2019a. doi: 10.2139/ssrn.3306006. URL https://ssrn.com/abstract=3306006.

Daniel Susser, Beate Roessler, and Helen F. Nissenbaum. Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2), 2019b. doi: 10.14763/2019.2.1410. URL https://policyreview.info/articles/analysis/technology-autonomy-and-manipulation. License: Creative Commons Attribution 3.0 Germany.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6199.

Steen Søndergaard. Cognitive warfare: Nato chief scientist research report. Research Report STO-OCS-001, NATO Science and Technology Organization (STO), Dec 2025.

Vivian Ta, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, and Alexia Loggarakis. User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of medical Internet research*, 22(3):e16235, 2020.

Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, 2008.

Nitasha Tiku and Sabrina Malhi. What is 'AI Psychosis' and How Can ChatGPT Affect Your Mental Health? *The Washington Post*, August 2025.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.

Jean M Twenge, Thomas E Joiner, Megan L Rogers, and Gabrielle N Martin. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among us adolescents after 2010 and links to increased new media screen time. *Clinical psychological science*, 6(1):3–17, 2018.

Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*, 6(1): 2056305120903408, 2020.

B. Vigliarolo. Darpa worried battlefield mixed reality vulnerable to 'cognitive attacks'. https://www.theregister.com/2023/10/12/darpa_worried_battlefield_mixed_reality/, Oct 12 2023.

Lee Vinsel. You're doing it wrong: Notes on criticism and technology hype. Medium (STS News), February 2021.

Rand Waltzman. The weaponization of information: The need for cognitive security. 2017. URL https://api.semanticscholar.org/CorpusID:39793649.

Jiashuo Wang, Kaitao Song, Chunpu Xu, Changhe Song, Yang Xiao, Dongsheng Li, Lili Qiu, and Wenjie Li. Enhancing user engagement in socially-driven dialogue through interactive llm alignments, 2025. URL https://arxiv.org/abs/2506.21497.

Kevin Wei, Patricia Paskov, Sunishchal Dev, Michael J Byun, Anka Reuel, Xavier Roberts-Gaal, Rachel Calcott, Evie Coxon, and Chinmay Deshpande. Position: Human baselines in model evaluations need rigor and transparency (with recommendations & reporting checklist). In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=gwhPvu97Gm.

Brenda K. Wiederhold. Inoculation theory in the digital age: Resilience against disinformation. *Cyberpsychology, Behavior, and Social Networking*, 2025. doi: 10.1177/21522715251399256.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Faxas Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, Krishan V Shenoy, and Jaimie M Henderson. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Wichayaporn Wongkamjan, Feng Gu, Yanze Wang, Ulf Hermjakob, Jonathan May, Brandon M Stewart, Jonathan K Kummerfeld, Denis Peskoff, and Jordan Lee Boyd-Graber. More victories, less cooperation: Assessing cicero's diplomacy play. *arXiv preprint arXiv:2406.04643*, 2024.

Jason Jia-Xi Wu. Beyond free markets and consumer autonomy: Rethinking consumer financial protection in the age of artificial intelligence. *NYU Journal of Intellectual Property & Entertainment Law*, 13(1):56–134, 2023. URL https://papers.ssrn.com/abstract=4566590. Published September 9, 2023; revised March 4, 2024.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. Collabllm: From passive responders to active collaborators. *arXiv preprint arXiv:2502.00640*, 2025.

Nick Yee, Jeremy N Bailenson, and Nicolas Ducheneaut. The proteus effect: Implications of transformed digital self-representation on online and offline behavior. *Communication Research*, 36(2):285–312, 2009.

Jinyan Yuan, Bangbang Yang, Keke Wang, Panwang Pan, Lin Ma, Xuehai Zhang, Xiao Liu, Zhaopeng Cui, and Yuewen Ma. Immersegen: Agent-guided immersive world generation with alpha-textured proxies, 2025. URL https://arxiv.org/abs/2506.14315.

Brahim Zarouali, Tom Dobber, Guy De Pauw, and Claes De Vreese. Using a personality-profiling algorithm to investigate political microtargeting: assessing the persuasion effects of personality-tailored ads on social media. *Communication Research*, 49(8):1066–1091, 2022.

Qingyu Zhang, Chunlei Xin, Xuanang Chen, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, Qing Ye, Qianlong Xie, and Xingxing Wang. Ai-salesman: Towards reliable large language model driven telemarketing, 2025. URL https://arxiv.org/abs/2511.12133.

Tao Zhou and Chunlei Zhang. Examining generative ai user addiction from a cac perspective. *Technology in Society*, 78:102653, 2024.

Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. The automated but risky game: Modeling and benchmarking agent-to-agent negotiations and transactions in consumer markets, 2025. URL https://arxiv.org/abs/2506.00073.

Tim Zindulka, Sven Goller, Daniela Fernandes, Robin Welsch, and Daniel Buschek. The ai memory gap: Users misremember what they created with ai or without. *arXiv preprint arXiv:2509.11851*, 2025.