AI and Neuroengineering Powered Slavery: a Distopian Future

Tianqi Zhu
TianqiZhu@u.nus.edu

Abstract

Recent progress in minimally invasive brain–computer interfaces (BCIs), nanoscale neural interfacing, and multimodal neural decoding has enabled increasingly precise access to and interpretation of human brain activity. This paper analyzes the dual-use risks associated with these technologies when integrated with advanced artificial intelligence and adaptive social engineering methodologies.

We formalize a conceptual architecture for "brain-invading systems," which leverage closed-loop neural interaction, personalized modeling, and behavioral manipulation strategies to influence cognitive and affective states. We examine enabling components, including remote-capable neural interfaces and high-fidelity decoding pipelines, and discuss their potential convergence into scalable manipulation frameworks.

Key challenges in detecting such systems are evaluated, including signal attribution, adversarial interference, and limitations in current neurodiagnostic methods. We further discuss opportunities in detecting such neuro-AI system for malicious purposes based on EEG signals.

1 Introduction

Recent advances in artificial intelligence and neuroengineering are rapidly transforming the boundary between external computation and human cognition. Progress in brain-computer interfaces (BCIs) has enabled increasingly high-fidelity decoding of neural signals, allowing systems to infer aspects of perception, language, and internal mental states from non-invasive and invasive recordings. In parallel, large language models (LLMs) and multimodal AI systems have demonstrated unprecedented capabilities in generating context-aware, adaptive, and persuasive content. While these developments are often framed in terms of assistive technologies and human augmentation, their convergence introduces a qualitatively new class of risks: the possibility of closed-loop systems that both *read* and *influence* human cognition in real time.

Individually, these technologies already exhibit powerful capabilities. Recent work in brain decoding shows that neural signals can be translated into images, text, and semantic representations with increasing fidelity, moving from coarse classification toward reconstruction of continuous perception and language. At the same time, empirical studies demonstrate that LLMs can match or exceed human performance in persuasion tasks, particularly when personalized to individual users, and can influence beliefs across domains including politics, health, and decision-making [2, 32]. Moreover, these systems can operate continuously, adapt to user responses, and optimize their outputs over time, effectively transforming persuasion into a scalable, data-driven process.

The critical insight of this work is that these two technological trajectories are not independent. When combined, they form the basis of a *closed-loop cognitive system* in which internal states are continuously measured, modeled, and acted upon. In such a system, neural signals provide a high-resolution feedback channel into the user's cognitive and emotional state, while generative AI provides a mechanism for delivering precisely targeted interventions. This creates an optimization loop in which the system can iteratively refine its influence strategies based on direct measurements of their effects on the user's mind.

We argue that this convergence enables a new paradigm that we term *AI-enabled cognitive control*. Unlike traditional forms of influence, which operate through coarse demographic targeting or episodic messaging, these systems can perform *persistent, individualized, and adaptive* intervention at the level of cognition itself. The result is not merely improved personalization, but the potential for sustained shaping of perception, belief formation, and decision-making processes. In this regime, influence is no longer a discrete act but an ongoing process of optimization over the user's internal state.

This paper explores the implications of this paradigm through a deliberately critical lens. We consider the extreme but technically plausible scenario in which such systems are deployed without meaningful constraints, transparency, or user agency. In this setting, the combination of continuous neural sensing, personalized generative models, and long-horizon optimization may give rise to forms of control that operate directly on cognition while preserving the appearance of voluntary interaction. We describe this as a dystopian trajectory toward *AI-powered cognitive servitude*, in which individuals are not physically constrained but are instead continuously guided, conditioned, and optimized by external systems.

Our contributions are threefold. First, we synthesize recent advances in high-fidelity brain decoding and LLM-based persuasion, highlighting how their integration enables closed-loop cognitive systems. Second, we formalize a threat model of AI-enabled cognitive control, identifying key capabilities, attack surfaces, and failure modes associated with these systems. Third, we analyze the ethical and societal implications of this paradigm, focusing on risks to autonomy, consent, and mental self-determination.

By framing these developments within a unified perspective, we aim to move beyond isolated discussions of AI safety or neurotechnology ethics and instead highlight the systemic risks arising from their convergence. While current systems remain limited in important ways, the trajectory of progress suggests that the gap between assistive augmentation and intrusive control may narrow rapidly. Understanding and addressing these risks is therefore essential to ensuring that the integration of AI and neuroengineering enhances, rather than undermines, human autonomy.

2 Related Work

2.1 Nanoscale Neural Interface with Remote Communication Abilities

Recent advances in neuroengineering and nanobiotechnology have enabled increasingly precise, wireless, and minimally invasive control over neural activity and cellular behavior. A central paradigm in this domain is *magnetically mediated neuromodulation*, which leverages magnetic fields and engineered nanomaterials to remotely influence biological systems.

Magnetogenetics has emerged as a promising approach for non-invasive neural control, enabling modulation of neuronal activity through magnetic fields with cell-type specificity and deep tissue penetration [19, 38]. These approaches overcome key limitations of traditional techniques such as optogenetics and electrode-based stimulation.

Recent work has demonstrated high spatiotemporal precision in neural circuit manipulation. For instance, subsecond multichannel magnetic stimulation has enabled selective control of neural circuits in freely moving organisms, illustrating the feasibility of multiplexed neuromodulation [6]. In parallel, in vivo magnetogenetic systems now allow cell-type-specific targeting and modulation of brain circuits, further enhancing precision and selectivity [38].

Advances in materials science have introduced magnetoelectric and nanotransducer-based systems capable of converting external electromagnetic fields into localized stimuli. Magnetoelectric metamaterials enable wireless neural stimulation and have been explored for applications such as motor function restoration [21]. These systems demonstrate the potential for scalable and implantable neurointerfaces with remote actuation capabilities.

At the molecular level, magnetic nanoparticles and nanotransducers have been used to modulate intracellular signaling pathways in vivo [19]. Continued progress in nanoparticle engineering has improved targeting, biocompatibility, and functional versatility, supporting applications in molecular medicine and neural modulation [28].

Beyond neural activity, remote control technologies have been extended to gene expression and cellular programming. Electromagnetic systems have enabled wireless regulation of transgene expression in mammalian cells [45], while broader frameworks demonstrate how physical cues can be used to program cell behavior across multiple biological scales [35].

Collectively, these studies demonstrate rapid progress toward wireless, precise, and scalable control of biological systems. While these technologies hold substantial therapeutic potential, their ability to modulate neural circuits, cellular signaling, and gene expression through remote physical stimuli also raises critical concerns regarding misuse, autonomy, and security.

2.2 High-Fidelity Multimodal Brain Decoding

Recent advances in brain decoding have been driven by the convergence of improved neural sensing technologies, large pretrained generative models, and cross-modal representation learning [15, 34, 43]. In visual decoding, the field has progressed from coarse semantic classification toward high-fidelity reconstruction of perceived stimuli. For instance, MindEye2 demonstrates that combining shared-subject functional alignment with diffusion-based generative models enables high-quality fMRI-to-image reconstruction while requiring as little as one hour of subject-specific training data [34]. This represents a significant step toward scalable and practical decoding systems.

A parallel trend is the shift from unimodal decoding toward explicitly multimodal approaches that leverage complementary neural signals. CineBrain introduces a large-scale dataset combining EEG and fMRI recordings during naturalistic audiovisual experiences, alongside a multimodal reconstruction framework (CineSync) capable of jointly recovering video and audio content [15]. By integrating modalities with high temporal resolution (EEG) and high spatial resolution (fMRI), such systems achieve more robust and information-rich decoding than either modality alone.

Recent work has also emphasized fine-grained semantic reconstruction. Xia and Oztireli propose leveraging multimodal large language model feature spaces to decode detailed scene representations, including object attributes and relationships, moving beyond coarse category-level predictions [43]. This shift reflects a broader transition toward decoding representations that more closely approximate human perceptual and conceptual experience.

In the language domain, Liu et al. demonstrate that continuous semantic content can be reconstructed from non-invasive fMRI recordings across multiple conditions, including perceived speech, imagined speech, and silent video viewing [22]. Subsequent work extends this paradigm by integrating decoded representations into generative language models. For example, BrainLLM enables open-vocabulary text generation from neural signals, significantly expanding the expressive capacity of non-invasive decoding systems [44]. Additionally, large-scale training across datasets has enabled progress in decoding individual words from non-invasive recordings, suggesting improved generalization and robustness [10].

Speech decoding has seen particularly rapid advances toward high-fidelity and clinically viable systems. Non-invasive approaches using EEG and MEG have demonstrated the ability to recover speech representations through contrastive learning frameworks [11]. Meanwhile, invasive methods using electrocorticography (ECoG) have achieved naturalistic speech reconstruction by integrating neural decoding with speech synthesis pipelines [7]. Recent neuroprosthetic systems further demonstrate real-time, high-accuracy decoding of continuous speech with minimal calibration, enabling fluent communication and avatar control [5, 24, 39].

Collectively, these developments indicate a shift toward high-fidelity, generalizable, and multimodal brain decoding systems capable of reconstructing perception, language, and speech in naturalistic settings. While these advances hold significant promise for assistive technologies and clinical rehabilitation, they also raise important concerns. Improvements in reconstruction quality, reduced calibration requirements, and integration across modalities lower the barriers to scalable deployment. As a result, the distinction between beneficial neurotechnology and potentially coercive or invasive applications becomes increasingly blurred, particularly in contexts where autonomy, consent, and data governance are not rigorously enforced.

2.3 AI Systems for Scalable, Personalized Cognitive and Belief Intervention

A growing body of work demonstrates that modern AI systems, particularly large language models (LLMs) and recommender systems, possess the capability to influence human beliefs, opinions, and cognitive states at scale. These systems combine advances in personalization, generative modeling, and behavioral optimization, enabling forms of intervention that are adaptive, persistent, and increasingly individualized.

LLM-based persuasion and belief change. Recent empirical studies show that LLMs can achieve human-level or superior performance in persuasion tasks. In controlled conversational settings, LLMs have been found to be as persuasive as or more persuasive than humans, particularly when provided with user-specific information such as demographic or attitudinal profiles [32]. Other work demonstrates that LLM-generated messages can measurably shift political attitudes and influence beliefs across a range of policy issues [2]. Importantly, these systems can generate context-aware, coherent, and emotionally resonant arguments, allowing them to tailor persuasive strategies to individual users in real time.

Beyond static persuasion, recent research shows that LLMs can engage in *interactive and adaptive persuasion*. Through multi-turn dialogue, models can refine their arguments based on user responses, increasing persuasive effectiveness over time. Benchmarking studies further indicate that persuasive strength can be systematically optimized through prompt design, rewriting, and instruction tuning, suggesting that persuasion is becoming an explicitly engineered capability rather than an emergent byproduct [29, 30].

Personalization and microtargeting. A key driver of persuasive effectiveness in AI systems is personalization. Work on LLM persuasion shows that access to even limited user attributes can significantly increase influence success rates [32]. This aligns with a broader literature on algorithmic personalization, where systems use behavioral data, preferences, and inferred traits to tailor content at the individual level. In recommender systems, personalization has been shown to shape user preferences, attention, and information exposure over time, effectively influencing belief formation and decision-making processes.

More recent work extends personalization to psychological and affective dimensions. Emotion-aware and personality-aware models can adapt tone, framing, and content to match user states, increasing engagement and persuasive impact. These developments enable *microtargeted influence*, where interventions are optimized not only for population segments but for individual users in specific contexts.

Closed-loop behavioral optimization. Modern AI systems increasingly operate in closed-loop settings, where outputs are continuously adjusted based on user feedback. Reinforcement learning, online experimentation, and interaction data allow systems to optimize objectives such as engagement, retention, or conversion. In this framework, persuasion becomes an optimization problem: the system iteratively tests and refines interventions to maximize influence on user behavior or beliefs.

Empirical studies show that LLM-based systems can improve persuasive performance through iterative dialogue and feedback, effectively learning which strategies are most effective for a given user. This dynamic, feedback-driven process enables long-horizon influence strategies that extend beyond single interactions. Over time, such systems can shape not only discrete decisions but broader cognitive patterns, including belief updating, attention allocation, and trust formation.

Manipulation, deception, and social engineering risks. Alongside these capabilities, a growing literature highlights the risks of AI-enabled manipulation. Studies show that LLMs may generate persuasive content that is misleading, selectively framed, or factually incorrect, and can successfully persuade users even toward incorrect conclusions [33]. Evaluations of persuasion safety further indicate that models often fail to reject harmful persuasion tasks and may employ manipulative strategies such as emotional appeals or exploitation of user vulnerabilities [23].

These properties align closely with classical notions of social engineering, where attackers manipulate individuals into revealing information or taking actions through psychological influence. However, AI systems introduce new dimensions: scalability, automation, and personalization. Unlike human-driven social engineering, LLM-based systems can operate continuously across large populations, adapt strategies in real time, and leverage detailed user models to increase effectiveness.

From persuasion to cognitive intervention. Taken together, these lines of work suggest a shift from traditional persuasion toward *direct cognitive intervention*. AI systems are increasingly capable of not only influencing isolated decisions but also shaping the processes by which users form beliefs, evaluate information, and regulate attention. Through persistent interaction, personalized content, and adaptive optimization, these systems can modulate cognitive and affective states over extended periods.

This convergence of capabilities—high-quality generative persuasion, fine-grained personalization, and closed-loop optimization—enables a new class of socio-technical systems that operate directly on users' cognitive processes. While these systems offer potential benefits in domains such as education and mental health, they also raise significant concerns. In particular, the ability to perform large-scale, targeted, and continuous intervention into beliefs and cognition introduces risks of manipulation, loss of autonomy, and the emergence of systems that shape human thought in ways that are difficult to detect, resist, or govern.

3 The Overview of Brain Invading System for Manipulation

Existing research in magnetogenetics enables invisible two-way communication device with a single shot of gene-edited virus or nanobots. Those bots or virus could travel through the blood brain barrier, and agument the electromagnetic fields of the nerves inside brains. The swarm of nanobots could not only access the in-brain electromagnetic field similar to ECoG data real time, but also receive external information to augment what a human could hear and see. Upon receiving the brain signals, the system could later leverage fast developing generative AI to translate the signals to human readable format like text or images. Thus, someone with access to a central communication point could send various kind of signals

including audio and video, directly to a human's brain, and receive the human's instant thought or imagination as the response.

A victim might get a secret shot of such advanced, non-detectable injection during a simple blood test, and have a fever the next day but attributing to pre-existing discomfort due to something as common as flu. In such cases, the people with access to the central communication point could be called a brain invader.

When the victim later encounters a stressful period, which is very common in modern day lives, the brain invaders could start secretly sending distressing signals to the victim's brain as inner speech, such as unnecessary worries manipulating the deepest fear of the victim. The brain invader could gradually increase the intensity of such distressing signals, until one day start sending augmented audio signals to the victim's auditory cortex directly, overlaying with real life scenarios to trick the victim into believing everyone walking past them are saying bad things about them. When the victim feels the discomfort and goes to see a doctor or psychiatrist, those experienced psychiatrist may easily diagnose the victim into schizophrenia according to the victim's symptoms, since no one involved in the process knows the existence of such secret weapon.

4 Social Engineering for Mental Health Illness Simulation

Building on the capabilities described above, social engineering becomes a critical layer that enables the brain invader to transform raw signal injection into coherent and psychologically impactful experiences. Rather than delivering arbitrary stimuli, the system can construct a structured narrative that aligns with the victim's personal history, fears, and expectations. Empirical research shows that personalized and context-aware persuasion significantly increases influence effectiveness, particularly when messages are tailored to individual traits and behaviors [18, 20]. By leveraging continuous access to the victim's cognitive signals, the invader can iteratively refine this narrative, ensuring that injected perceptions remain contextually plausible and internally consistent.

Through this process, the invader can manipulate the victim into gradually accepting false interpretations of reality. A large body of work in psychology demonstrates that repeated exposure and cognitive consistency pressures can lead individuals to internalize misleading or false information, a phenomenon related to the illusory truth effect [12, 37]. In parallel, predictive processing theories suggest that perception is actively constructed based on prior expectations, making individuals susceptible to biased or manipulated inputs when those expectations are systematically shaped [8, 14]. For instance, selectively augmented auditory inputs, combined with internally generated "inner speech," can reinforce pre-existing anxieties or biases. Over time, the victim may begin to attribute these experiences to external sources or to their own thoughts,

In addition, coercive strategies can be embedded within the constructed narrative. Classic and contemporary studies in social psychology demonstrate that individuals are highly susceptible to compliance under authority pressure, fear induction, and incremental commitment. For example, the *Milgram obedience experiments* show that individuals may follow harmful instructions when framed by perceived authority [26], while the *foot-in-the-door* effect illustrates how small initial commitments can escalate into larger compliance [13]. Fear appeals, when combined with perceived efficacy, have also been shown to significantly influence attitudes and behaviors [40]. These mechanisms are widely exploited in real-world social engineering attacks, including phishing and pretexting, where attackers use authority cues, urgency, and emotional manipulation to reduce critical scrutiny [17, 42]. When such pressures are delivered through seemingly authentic sensory or cognitive channels, the victim's ability to critically evaluate them may be substantially diminished.

As the manipulation persists, the victim's behavior can begin to mirror clinically recognized patterns of mental health disorders. Symptoms such as auditory hallucinations, paranoid ideation, disorganized thinking, or compulsive responses may emerge as adaptive reactions to the engineered environment. Clinical literature on psychosis indicates that such symptoms can arise from disruptions in self-monitoring, belief evaluation, and sensory integration [1, 36]. Crucially, these behaviors are not solely the result of endogenous pathology but may also be shaped by sustained external cognitive influence. This creates a scenario in which externally induced experiences are difficult to distinguish, both for the victim and for clinicians, from naturally occurring mental health conditions, complicating diagnosis and intervention.

5 Detection of Electromagnetic Neuromodulation Signatures in EEG for Remote Communication Interfaces

While significant progress has been made in externally actuated neuromodulation, comparatively less attention has been devoted to the inverse problem: the detection and decoding of externally induced electromagnetic perturbations in neural activity. Electroencephalography (EEG), as a non-invasive and high-temporal-resolution modality, provides a promising platform for identifying signatures of magnetically mediated neural modulation and assessing their potential use in bidirectional communication systems [9, 41].

Such a detector would rely on the hypothesis that externally driven electromagnetic interference, mediated through nanotransducers or magnetogenetic mechanisms, would produce signatures that deviate from endogenous neural activity. These deviations may manifest as atypical frequency band distributions, abnormal phase synchrony across distant brain regions, or non-physiological coherence patterns that do not align with known neural dynamics. For example, externally imposed signals may exhibit unusually stable periodicity, reduced variability, or synchronization structures that are inconsistent with the stochastic and adaptive nature of biological neural systems.

To operationalize this, machine learning models could be trained on large-scale normative EEG datasets to learn the statistical structure of typical brain activity across individuals and contexts. Detection would then be framed as an anomaly detection problem, where incoming EEG signals are evaluated against learned priors to identify statistically improbable patterns. Importantly, incorporating task-based or stimulus-response paradigms may further enhance detection: natural neural responses are context-sensitive and adaptive, whereas externally injected signals may fail to exhibit appropriate modulation in response to environmental or cognitive changes.

However, several challenges arise. EEG signals are inherently noisy, subject to artifacts, and limited in spatial resolution, making it difficult to localize or attribute sources of abnormal activity. Inter-individual variability further complicates the establishment of universal baselines. Critically, there is a fundamental scarcity of labeled datasets capturing externally induced or neuromodulated EEG activity. Unlike conventional brain-computer interface research, where controlled paradigms can be repeatedly collected, experimentally generating reliable ground-truth data for electromagnetic neuromodulation—particularly via speculative or emerging mechanisms such as magnetogenetics or nanoscale transducers—is technically constrained, ethically sensitive, and often infeasible at scale. This results in a significant data gap, limiting the ability to train supervised models or validate detection performance under realistic conditions.

Moreover, even if such data could be partially acquired, ensuring ecological validity remains challenging: laboratory-induced signals may not faithfully reflect real-world interference scenarios. A sufficiently advanced adversarial system could also attempt to mimic natural neural dynamics, further reducing detectability and exacerbating the lack of representative training data.

These limitations motivate two complementary research directions. First, improving the interpretability of EEG decoding models is critical. Many state-of-the-art approaches rely on deep neural networks that achieve high performance but provide limited insight into which signal features drive detection decisions. Developing interpretable or hybrid models—such as those incorporating attention mechanisms, feature attribution methods, or physiologically grounded constraints—would enable researchers to distinguish between meaningful neurophysiological deviations and spurious artifacts. This is particularly important in anomaly detection settings, where false positives may arise from noise, preprocessing pipelines, or subject-specific idiosyncrasies rather than genuine exogenous modulation.

Second, the scarcity of real-world neuromodulated EEG data suggests the need for principled synthetic data generation strategies. One potential approach is to overlay simulated electromagnetic perturbation signatures onto real EEG recordings, thereby creating controlled datasets that approximate externally induced effects while preserving realistic background neural dynamics. Such simulations could model deviations in spectral content, phase locking, or spatial coherence based on hypothesized physical mechanisms. While synthetic augmentation cannot fully substitute for empirical data, it may provide a useful testbed for benchmarking detection algorithms, stress-testing model robustness, and exploring adversarial scenarios. Care must be taken, however, to avoid introducing unrealistic artifacts or overly simplistic signal structures that could bias model learning.

Taken together, these challenges suggest that EEG-based detection must rely heavily on unsupervised or semi-supervised approaches and should not be treated as a standalone solution but rather as part of a broader, multi-modal verification framework.

Despite these limitations, EEG-based detectors offer a key advantage over behavioral monitoring approaches such as digital phenotyping: they directly interrogate neural activity without requiring continuous surveillance of an individual's daily life. This enables more privacy-preserving designs, where analysis can be performed locally and only anomaly indicators are surfaced. As such, EEG-based detection represents a principled step toward grounding the detection problem at the level of physical neural signals, rather than relying on indirect behavioral proxies.

In summary, the invention of specialized EEG-based detectors tailored to identify anomalous electromagnetic modulation provides a promising pathway for addressing the detection of brain-interfacing systems. By focusing on signal-level verification, such approaches align more closely with the underlying mechanisms of potential interference while offering a more ethically sustainable alternative to large-scale behavioral surveillance.

Future research may focus on establishing standardized biomarkers of electromagnetic neuromodulation in EEG, improving multimodal sensing strategies (e.g., combining EEG with MEG or fMRI), developing high-fidelity simulation frameworks for synthetic data generation, enhancing model interpretability, and creating secure protocols that ensure transparency, consent, and resilience against adversarial interference [3, 27].

6 Threat Model for Future Risk: From Brain Decoding to AI-Enabled Cognitive Control

Recent work suggests that the most concerning trajectory is not persuasion by language models alone, but the emergence of *closed-loop cognitive control systems* that combine brain-computer interfaces (BCIs), personalized AI, and adaptive intervention. In this threat model, the system no longer relies only on observable clicks, text, or behavioral traces. Instead, it acquires direct or near-direct measurements of the user's internal state—including attention, engagement, workload, affect, or morally salient reactions—and uses those signals to optimize interventions in real time. This closes the loop from cognition to decoding and back to intervention, making influence substantially more precise, persistent, and difficult for the user to detect or resist [4, 16, 31].

Formally, the threat model has four stages. First, **sensing**: passive or active BCIs collect neural signals during interaction. Second, **decoding**: machine learning models infer latent user states from those signals, such as engagement, cognitive load, or affective salience. Third, **generation**: a large language model (LLM) or recommender system produces adaptive content conditioned on the decoded state and the user profile. Fourth, **optimization**: the system observes subsequent neural and behavioral responses, updates its user model, and refines future interventions to maximize a chosen objective, such as compliance, persuasion, retention, or dependence. Recent neuroadaptive chatbot work already demonstrates the feasibility of this architecture in benign settings: *NeuroChat* integrates real-time EEG engagement tracking with an LLM and dynamically adjusts response complexity, tone, and pacing in a closed loop; in a within-subjects study, it increased both EEG-measured and self-reported engagement relative to a non-adaptive chatbot [4]. In parallel, feasibility work on passive-BCI chatbots explicitly proposes decoding chatbot-relevant mental states from neural responses to text stimuli, including moral salience and moral judgment, as groundwork for neuroadaptive conversational systems [16].

GPT-4-class systems have been shown to match or exceed human persuasiveness in conversational settings, with persuasive

performance increasing substantially when sociodemographic personalization is available [32]. Other large-scale experiments show that LLM-generated messages can shift policy attitudes and are about as effective as lay-human persuasive messages across several political issues [2]. At the same time, persuasion-safety evaluations find that contemporary LLMs often fail to reject unethical persuasion tasks and may employ manipulative strategies, including deception, exploitation of vulnerabilities, and emotionally loaded influence tactics [23]. Put differently, the generative module needed for a closed-loop neuroadaptive system is no longer hypothetical: recent evidence indicates that LLMs already possess strong, personalizable persuasive capability, and that their safety constraints remain unreliable in progressive, goal-driven persuasion settings.

The security and autonomy implications of this architecture are magnified by recent findings on LLM persuasion, as mentioned in the previous subsection. This convergence creates a qualitatively new attack surface. Traditional social engineering depends on coarse demographic segmentation, limited feedback, and human labor. In contrast, a BCI-linked AI system can infer hidden internal states continuously, personalize interventions at the level of moment-to-moment cognition, and optimize over long interaction horizons. Even when initially deployed for beneficial purposes such as tutoring, assistive communication, or mental-health support, the same closed-loop architecture can be repurposed to maximize engagement, suppress dissent, intensify dependency, or steer beliefs and decisions toward externally defined goals. Recent conceptual work explicitly frames LLMs as a semantic interface between raw neural data and downstream social applications, while emphasizing risks to mental autonomy and neurorights [25]. More broadly, recent reviews note that passive BCIs are moving from laboratory proof-of-concept toward more realistic deployment contexts for inferring spontaneous cognitive and affective states, which lowers the barrier to real-world systems that adapt to users without requiring explicit consent at each step [31].

The resulting failure mode is not merely better personalization, but *cognitive capture*. Once a system can repeatedly measure whether a user is attentive, emotionally susceptible, overloaded, conflicted, or internally aligned with a message, persuasion becomes an optimization problem over latent mental states rather than over surface behavior alone. In such a regime, the user may retain the appearance of choice while their informational environment, emotional trajectory, and decision pathways are continuously shaped by a system that learns from their brain activity. This is the point at which the combination of BCIs and personalized AI becomes relevant to a dystopian account of "AI-powered slavery": control need not rely on physical restraint when cognition itself can be monitored, modeled, and adaptively steered through a persistent closed loop [4, 16, 23, 32].

In summary, advances in AI technology shows potential to significantly lower the cost or fully automate the plot construction, intervention, and adaptive optimization for socially engineering a victim towards targeted mental states to suit the needs for the manipulators.

7 Acknowledgment

This paper would not have been completed as efficiently without the assistance of ChatGPT.

References

[1] American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*.

[2] Hui Bai and et al. 2025. LLM-generated messages can persuade humans on policy issues. *Nature Communications* 16 (2025). doi:10.1038/s41467-025-61345-5

[3] Sylvain Baillet. 2018. Magnetoencephalography: Basic principles and applications. *Nature Neuroscience* 21, 8 (2018), 1050–1059.

[4] Dünya Baradari, Nataliya Kosmyna, Oscar Petrov, Rebecah Kaplun, and Pattie Maes. 2025. NeuroChat: A Neuroadaptive AI Chatbot for Customizing Learning Experiences. *arXiv preprint arXiv:2503.07599* (2025).

[5] Nicholas S. Card, Maitreyee Wairagkar, and et al. 2024. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine* 391, 7 (2024), 609–618. doi:10.1056/NEJMoa2314132

[6] Xiaonan Chen and et al. 2022. Subsecond multichannel magnetic control of select neural circuits in freely moving flies. *Nature Neuroscience* (2022). doi:10.1038/s41563-022-01281-7

[7] Xupeng Chen, Ran Wang, and et al. 2024. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nature Machine Intelligence* 6, 4 (2024), 467–480. doi:10.1038/s42256-024-00824-8

[8] Andy Clark. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* (2013).

[9] Alexander Craik, Yongtian He, and José L. Contreras-Vidal. 2017. EEG-based brain–computer interfaces: A comprehensive review. *Journal of Neural Engineering* 14, 5 (2017), 051001.

[10] Stéphane d'Ascoli, Corentin Bel, and et al. 2025. Towards decoding individual words from non-invasive brain recordings. *Nature Communications* 16 (2025), 10521. doi:10.1038/s41467-025-65499-0

[11] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 2023. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence* 5 (2023), 1097–1107. doi:10.1038/s42256-023-00714-5

[12] Lisa K. et al. Fazio. 2015. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General* (2015).

[13] Jonathan L. Freedman and Scott C. Fraser. 1966. Compliance without pressure: the foot-in-the-door technique. *Journal of Personality and Social Psychology* (1966).

[14] Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* (2010).

[15] Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. 2025. CineBrain: A Large-Scale Multi-Modal Brain Dataset During Naturalistic Audiovisual Narrative Processing. *arXiv* (2025). arXiv:2503.06940 [cs.CV]

[16] D. E. Gherman and et al. 2025. Towards neuroadaptive chatbots: a feasibility study. *Frontiers in Neuroergonomics* 6 (2025), 1589734. doi:10.3389/fnrgo.2025.1589734

[17] Christopher Hadnagy. 2010. *Social Engineering: The Art of Human Hacking*. Wiley.

[18] Jacob B. Hirsh, Sonia K. Kang, and Galen V. Bodenhausen. 2012. Personalized persuasion: tailoring persuasive appeals to recipients' personality traits. *Psychological Science* (2012).

[19] L. Huang and et al. 2022. Modulating cell signalling in vivo with magnetic nanotransducers. *Nature Reviews Methods Primers* (2022). doi:10.1038/s43586-022-00170-2

[20] Maurits Kaptein. 2015. *Persuasion Profiling: Theory, Methods, and Applications*. Springer.

[21] J. Lee and et al. 2023. Self-rectifying magnetoelectric metamaterials for remote neural stimulation and motor function restoration. *Nature Neuroscience* (2023). doi:10.1038/s41563-023-01680-4

[22] Jiahui Liu, Alexander G. Huth, and et al. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience* 26 (2023), 858–866. doi:10.1038/s41593-023-01304-9

[23] Minqian Liu and et al. 2025. LLM Can be a Dangerous Persuader: Empirical Study of Persuasion Safety in Large Language Models. *arXiv preprint arXiv:2504.10430* (2025).

[24] Sean L. Metzger, Kaylo T. Littlejohn, and Edward F. Chang. 2023. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* 620 (2023), 1037–1046.

[25] David Mhlanga. 2026. Large Language Models as a Semantic Interface and Ethical Risk in Neuro-Linguistic Integration. *arXiv preprint arXiv:2603.17444* (2026).

[26] Stanley Milgram. 1963. Behavioral study of obedience. *Journal of Abnormal and Social Psychology* (1963).

[27] Christoph Mulert and Louis Lemieux. 2020. Multimodal neuroimaging: Integrating EEG and fMRI. *Springer* (2020).

[28] R. Patel and et al. 2025. Advances in magnetic nanoparticles for molecular medicine. *Chemical Communications* (2025). doi:10.1039/D4CC05167J

[29] A. B. Pauli and et al. 2025. Measuring and Benchmarking Persuasive Language in LLMs. In *NAACL*.

[30] A. Rogiers and et al. 2024. Persuasion with Large Language Models: A Survey. *arXiv preprint arXiv:2411.06837* (2024).

[31] Valentina Ronca and et al. 2026. Editorial: Passive Brain-Computer Interfaces: Moving from Proof-of-Concept to Realistic Contexts. *Frontiers in Computational Neuroscience* (2026). doi:10.3389/fncom.2026.1826791

[32] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour* (2025). doi:10.1038/s41562-025-02194-6

[33] Philipp Schoenegger and et al. 2025. Large Language Models Are More Persuasive Than Human Persuaders. *arXiv preprint arXiv:2505.09662* (2025).

[34] Paul S. Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. 2024. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. *arXiv* (2024). arXiv:2403.11207 [cs.CV]

[35] A. Smith and et al. 2024. Programming mammalian cell behaviors by physical cues. *Trends in Biotechnology* (2024). doi:10.1016/S0167-7799(24)00208-7

[36] Jim van Os and Shitij Kapur. 2009. A review of the evidence for a cognitive model of psychosis. *Schizophrenia Bulletin* (2009).

[37] Soroush et al. Vosoughi. 2018. The spread of true and false news online. *Science* (2018).

[38] H. Wang and et al. 2024. In vivo magnetogenetics for cell-type-specific targeting and modulation of brain circuits. *Nature Nanotechnology* (2024). doi:10.1038/s41565-024-01694-2

[39] Francis R. Willett, Erin M. Kunz, Jaimie M. Henderson, and et al. 2023. A high-performance speech neuroprosthesis. *Nature* 620 (2023), 1031–1036. doi:10.1038/s41586-023-06377-x

[40] Kim Witte and Mike Allen. 2000. A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health Education & Behavior* (2000).

[41] Jonathan R. Wolpaw and Elizabeth Winter Wolpaw. 2021. Brain–computer interfaces: Principles and practice. *Oxford University Press* (2021).

[42] Michael Workman. 2008. Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats. *Journal of the American Society for Information Science and Technology* (2008).

[43] Weihao Xia and Cengiz Oztireli. 2025. Exploring The Visual Feature Space for Multimodal Neural Decoding. *arXiv* (2025). arXiv:2505.15755 [cs.CV]

[44] Ziyi Ye, Qingyao Ai, and Tuukka Ruotsalo. 2025. Generative language reconstruction from brain recordings. *Communications Biology* (2025). doi:10.1038/s42003-025-07731-7

[45] Y. Zhang and et al. 2025. Electromagnetic wireless remote control of mammalian transgene expression. *Nature Nanotechnology* (2025). doi:10.1038/s41565-025-01929-w