Unseen Hands: How Artificial Neural Architectures Influence Human Emotions and Decision-Making

Jay Viramgami
Department of Artificial Intelligence and Machine Learning
L.D. College of Engineering
Ahmedabad, India
22aijay075@ldce.ac.in

***Abstract*—This paper investigates the profound and often invisible influence of artificial neural architectures on human psychology. As artificial intelligence (AI) systems become integral to our cognitive ecosystem, their outputs—shaped by underlying architectural designs—directly modulate human emotions, decision-making processes, and neural plasticity. We synthesize foundational theories from psychology (Dual-Process Theory, Somatic Marker Hypothesis, Appraisal-Tendency Framework) and neuroscience with a technical analysis of key AI architectures, including Transformer models, Generative Adversarial Networks (GANs), and Reinforcement Learning (RL) agents. Through a review of experimental findings and the proposal of novel research designs, we demonstrate how variations in AI outputs can amplify cognitive biases, induce specific emotional states, and foster cognitive dependency. We present a comparative analysis of different AI models, such as Large Language Models (LLMs) and recommendation systems, to illustrate how architectural choices lead to distinct psychological impacts. The discussion culminates in an examination of the ethical and societal implications, including the potential for digital manipulation, the erosion of cognitive autonomy, and the long-term reshaping of human brain function. We conclude by advocating for a human-centered approach to AI development that prioritizes psychological well-being and cognitive sovereignty.**

***Index Terms*—Artificial Intelligence, Affective Computing, Human-AI Interaction, Cognitive Psychology, Neuroplasticity, AI Ethics**

I. Introduction: The New Cognitive Ecosystem

A. The Emergence of AI as a Cognitive Partner

The integration of artificial intelligence into the fabric of daily life marks a paradigm shift in the human cognitive experience. Far from being passive instruments, modern AI systems function as active, interactive agents within our cognitive environment, capable of summarizing information, generating novel content, engaging in dialogue, and making choices [1], [2]. This transition reframes the relationship between humans and technology, moving beyond the traditional model of Human-Computer Interaction (HCI), which studies how users interact with a system to complete a task, toward a model of Human-Cognition Symbiosis. In this symbiotic relationship, AI is not merely a tool to be used, but a system that actively participates in and reshapes the very processes of human thought, emotion, and decision-making. Just as the advent of literacy fundamentally altered the structure of human memory and abstract reasoning, the pervasive influence of AI is poised to reconfigure our internal psychological landscape in ways that are only beginning to be understood.

The central thesis of this paper is that the architectural choices made during the design of an AI model are not merely technical specifications but are direct, causal factors in shaping human psychological responses. The ”unseen hands” of these architectures—the mathematical logic of a self-attention mechanism, the optimization objective of a reinforcement learning agent, or the adversarial dynamic of a generative network—exert a potent and predictable influence on the user. This influence is not superficial; it extends to the core of human cognition, affecting everything from immediate emotional reactions to long-term belief formation and even the physical structure of the brain through neuroplasticity [3]. Research on children, for example, reveals that aggressive behavior toward AI systems can increase aggressive behavior toward humans, indicating that interactions with AI are not psychologically isolated but actively reshape social cognition [4]. This deep integration necessitates a shift in focus from questions of usability and efficiency to a more profound inquiry into the long-term alteration of human cognitive and emotional architecture.

B. Significance and Urgency

The urgency of this investigation is underscored by the rapid and widespread deployment of sophisticated AI systems into high-stakes domains. AI now plays a significant role in mental health support, where conversational agents offer therapy and companionship [5], [6]; in finance, where algorithms provide investment advice; and in education, where personalized platforms guide the learning process. In each of these areas, the outputs of AI models are directly influencing human emotions and guiding critical life decisions. Yet, this influence is often exerted without the user’s full awareness of the underlying mechanisms and without comprehensive regulatory oversight to ensure psychological safety. The potential for AI-driven filter bubbles to amplify confirmation bias, for emotionally resonant chatbots to foster unhealthy dependency, and for biased algorithms to perpetuate societal inequities represents a significant challenge to individual autonomy and social well-being. Understanding the intricate connections between AI architecture and human psychology is therefore not merely an academic exercise but a critical prerequisite for developing

responsible, ethical, and truly beneficial artificial intelligence [7].

C. Paper Structure and Interdisciplinary Approach

To construct a holistic understanding of this complex phenomenon, this paper adopts a deeply interdisciplinary approach, weaving together theoretical frameworks and empirical evidence from psychology, neuroscience, computer science, and ethics. The analysis will proceed as follows: Section II establishes the theoretical foundations by reviewing key models of human emotion and decision-making and theories of human-AI interaction. Section III provides a technical deconstruction of the primary neural architectures involved, explaining the mechanics through which they process and generate influential content. Section IV synthesizes existing experimental evidence and proposes novel research designs to empirically measure AI's psychological impact. Section V conducts a comparative analysis of different AI systems, such as Large Language Models and recommendation systems, to highlight how architectural variations lead to distinct psychological outcomes. Section VI discusses the profound long-term implications of this human-AI symbiosis, focusing on neuroplasticity and the pressing ethical and societal challenges that emerge. Finally, the conclusion summarizes the paper's key findings and advocates for a future research agenda and a design philosophy centered on preserving and enhancing human cognitive and emotional well-being. This integrated approach aims to move beyond siloed analyses to provide a systems-level view of one of the most significant cognitive transformations in human history.

II. Theoretical Foundations of Emotion, Cognition, and Interaction

To comprehend how artificial neural architectures influence human psychology, it is essential to first understand the established mechanisms of human emotion and decision-making. This section reviews foundational theories from psychology and neuroscience, connecting them to the emerging field of human-AI interaction. These frameworks provide the vocabulary and conceptual models needed to explain why and how AI-generated outputs can exert such a powerful and predictable influence on human behavior.

A. Psychological and Neuroscientific Models of Decision-Making

Human decision-making is not a purely rational process but a complex interplay of rapid intuition, deliberate reasoning, and potent emotional signals. Modern AI systems, through their design and optimization, often engage these processes in specific and targeted ways.

1) Dual-Process Theory in the Age of AI: Dual-Process Theory posits that human thought arises from two distinct cognitive systems [8]. System 1 (or Type 1) processing is fast, automatic, intuitive, and often emotionally driven. It relies on heuristics—mental shortcuts—to make quick judgments with minimal cognitive effort [9], [10]. System 2 (or Type 2) processing, in contrast, is slow, deliberate, analytical, and requires conscious effort and working memory [11]. While System 2 is responsible for complex reasoning and logical analysis, the vast majority of our daily decisions are guided by the efficiency of System 1 [12].

This framework is critically relevant to human-AI interaction. Many contemporary AI systems, particularly recommendation algorithms and social media content feeds, are architecturally optimized to engage and exploit System 1 processes. By presenting users with a continuous stream of emotionally charged, easily digestible content, these systems can trigger intuitive reactions and guide behavior while bypassing the reflective scrutiny of System 2. For example, the availability heuristic, where judgments are based on the ease with which examples come to mind, can be powerfully manipulated by an AI that repeatedly shows users vivid or emotionally charged content, skewing their perception of risk or prevalence [13]. This dynamic suggests that AI can act as a powerful "nudging" force, shaping choices by appealing directly to the heuristic-driven, low-effort nature of our intuitive minds.

2) The Somatic Marker Hypothesis and AI-Elicited Feelings: Antonio Damasio's Somatic Marker Hypothesis (SMH) challenges the notion that emotion is detrimental to rational decision-making [14]. Instead, it proposes that emotional signals are essential guides, particularly in complex and uncertain situations [15]. According to the SMH, when we encounter a situation, our brain retrieves memories of similar past experiences and reactivates the associated emotional states. These emotions manifest as physiological changes in the body—a rapid heartbeat, a knot in the stomach, a feeling of warmth—which Damasio termed "somatic markers" [13]. These "gut feelings" act as biasing signals, marking potential choices as either advantageous or disadvantageous, thereby narrowing down the options for more deliberate analysis [16].

Neuroscientifically, this process is believed to involve the ventromedial prefrontal cortex (VMPFC), which integrates these emotional signals, and the amygdala, which is crucial for processing emotional stimuli and forming these associations [9]. Patients with damage to the VMPFC, while retaining their intellectual capacities, often show severe impairments in real-life decision-making because they lack access to these emotional guiding signals [15].

This hypothesis provides a powerful framework for understanding AI's influence. AI systems, through the content they generate or recommend, can become potent external triggers for these internal somatic states. A suspenseful movie trailer recommended by a streaming service, an outrage-inducing news headline curated by a social media algorithm, or an empathetic response from a chatbot can all evoke distinct physiological responses. These AI-elicited somatic markers can then non-consciously bias subsequent decisions, from consumer choices to social judgments. The AI, in this sense, does not merely present information; it actively induces the physiological states that our brains have evolved to use as a primary input for decision-making. It functions as a "Somatic Marker Generator," an externalized component of a deeply

internal cognitive loop, hijacking a fundamental mechanism of autonomous choice.

3) The Appraisal-Tendency Framework (ATF): While the SMH explains the role of general emotional feeling, the Appraisal-Tendency Framework (ATF), developed by Lerner and Keltner, provides a more granular model for how specific emotions shape judgment and choice [17]. The ATF posits that each distinct emotion is associated with a unique set of cognitive appraisals—evaluations of a situation along dimensions such as certainty, control, and responsibility [18]. For example, fear is characterized by appraisals of low certainty and low personal control, whereas anger is characterized by appraisals of high certainty and high control, even though both are negatively valenced [19]. Crucially, the ATF proposes that an emotion, once activated, triggers an "appraisal tendency"—a predisposition to perceive and interpret subsequent, unrelated situations through the lens of that emotion's characteristic appraisals [17]. A person made to feel fearful will tend to perceive new situations as uncertain and risky, leading to risk-averse choices. In contrast, a person made to feel angry will perceive new situations as more certain and controllable, leading to more optimistic and risk-seeking choices [19].

The implications for AI are profound. AI systems can be designed to generate content that reliably elicits specific emotions. An AI curating a news feed can select articles that induce anger, thereby priming users to make more optimistic and risk-seeking decisions in an unrelated context, such as financial trading. Conversely, by surfacing content that evokes fear, it could prime more cautious, risk-averse behavior. The AI thus acts as an "Appraisal Setter," curating an informational environment that primes a specific cognitive lens through which the user views the world. This mechanism allows for a highly targeted and predictable form of influence that operates by shaping the very cognitive predispositions that underlie judgment.

4) Synthesizing Frameworks: The Emotion-Imbued Choice Model: The emotion-imbued choice model serves as a unifying framework that formally integrates the inputs from traditional rational choice theory with the potent, pervasive, and predictable influence of emotions [20]. It acknowledges that decisions are not made in a vacuum of pure reason but are deeply intertwined with our affective states [21]. This model provides a comprehensive lens through which to analyze human-AI decision-making, accounting for both the deliberative, goal-oriented aspects of choice and the powerful, often non-conscious, drivers of emotion that AI systems are increasingly adept at manipulating.

B. Theories of Human-AI Relational Dynamics

The influence of AI is not solely transactional; it is also relational. The way humans perceive, trust, and form bonds with AI systems fundamentally shapes how they respond to the information and emotional cues these systems provide.

1) Trust, Anthropomorphism, and Mental Models: Trust is a cornerstone of effective human-AI interaction, influencing user acceptance, reliance, and satisfaction [22]. Trust in an AI system is multifaceted, shaped by its perceived reliability and performance (ability), its transparency and fairness (integrity), and the user's belief that it is acting in their best interest (benevolence) [23]. Transparency is particularly crucial; users are more likely to trust systems that provide clear explanations for their decisions [22].

This process of trust formation is heavily influenced by anthropomorphism, the tendency to attribute human-like qualities, intentions, and emotions to non-human agents [22]. AI systems with human-like features, such as conversational chatbots or virtual avatars, often evoke stronger emotional and social responses [24]. This can be a double-edged sword: while it can foster engagement and social bonding, it can also lead to misplaced trust and vulnerability to manipulation [22].

The user's mental model of the AI—their internal representation of how the system works and what its intentions are—is a key determinant of the interaction's outcome. Strikingly, research has shown that simply priming users with a belief about an AI's motive (e.g., telling them it is a "caring" agent) can significantly increase their perception of its trustworthiness and empathy, even when the underlying AI system remains unchanged [25]. This highlights that the framing and introduction of an AI system are as important as its technical capabilities in shaping the human psychological response.

2) The Machine-Integrated Relational Adaptation (MIRA) Model: The Machine-Integrated Relational Adaptation (MIRA) model offers a sophisticated, transdisciplinary framework for understanding AI's evolving role in our social lives [26]. MIRA moves beyond simple interaction to conceptualize AI in two distinct relational roles:

1) **Relational Partner:** The AI as a direct interaction companion, such as a therapeutic chatbot or a social robot.
2) **Relational Mediator:** The AI as an intermediary that shapes human-to-human communication, such as a recommendation algorithm suggesting topics of conversation or a language model editing an email.

Central to MIRA are four principles that describe how AI fosters social adaptation: linguistic reciprocity, psychological proximity, interpersonal trust, and relational substitution versus enhancement [26]. By integrating established psychological theories like attachment theory and social exchange theory, MIRA provides a structured approach for analyzing how adaptive AI language and behavior can elicit emotional investment, simulate mutual understanding, and, in some cases, even supplant genuine human interaction [26]. This model provides a critical lens for examining the long-term social and emotional consequences of embedding AI deeply within human ecosystems.

III. Neural Architectures and the Mechanics of Influence

To fully grasp how AI systems exert psychological influence, it is necessary to move beyond theoretical models and examine the technical "unseen hands" themselves: the neural architectures that process and generate emotionally and

cognitively salient content. The choice of architecture is not a neutral design decision; it fundamentally determines the types of patterns an AI can learn, the kinds of outputs it can produce, and consequently, the specific psychological levers it can pull. This path dependence means that the ethical and psychological profile of an AI system is, to a large extent, predetermined by its core design.

A. Architectures for Affective Understanding

Before an AI can influence emotion, it must first be able to recognize it. Affective Computing is the interdisciplinary field dedicated to creating systems that can recognize, interpret, process, and simulate human affects [27]. This field categorizes its primary tasks into Affective Understanding (AU), the recognition and interpretation of emotion, and Affective Generation (AG), the creation of emotionally resonant content [28]. The evolution of AU architectures reveals a progression toward increasingly nuanced and context-aware models of human emotion.

1) Processing Emotional Text: The journey from raw text to emotional insight involves several key architectural components. Early methods relied on emotion dictionaries and manual feature extraction, but these have been largely superseded by deep learning models that can automatically learn rich representations of language [29].

- **Word Embeddings:** The foundational step in modern Natural Language Processing (NLP) is the conversion of words into dense numerical vectors, or embeddings. Unlike simple one-hot encodings, these vectors capture semantic relationships, such that words with similar meanings are located closer to each other in the vector space [30]. Models like Word2Vec and GloVe generate static embeddings, while more advanced models like ELMO (Embeddings from Language Models) produce contextualized embeddings, meaning the vector for a word like ”bank” will differ depending on whether it appears in a financial or geographical context [31]. This ability to capture context is crucial for accurately interpreting emotional nuance.
- **Recurrent and Convolutional Neural Networks (RNNs & CNNs):** Once text is converted to embeddings, different architectures can be used to extract meaning. RNNs, particularly variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), are designed to process sequential data. They maintain an internal ”memory” state, allowing them to capture long-range dependencies and understand the order of words in a sentence, which is vital for contextual understanding [29]. Bidirectional LSTMs process text in both forward and backward directions, further enriching the contextual representation [30]. CNNs, traditionally used for image processing, can be effectively applied to text by treating sentences as a 1D grid. They use filters (kernels) of different sizes to slide over the text and detect key local patterns or features (n-grams), such as emotionally charged phrases, regardless of their position in the sentence [29].
- **Ensemble Models:** State-of-the-art emotion classification often employs ensemble or hybrid architectures that combine these approaches. For instance, a model might use CNN layers to extract salient, time-invariant features from the text and then use these features to initialize the hidden state of an LSTM layer. This allows the model to benefit from both the local feature detection of CNNs and the sequential context modeling of RNNs, leading to a more robust understanding of emotional content [31].

2) The Transformer Architecture and Self-Attention: The introduction of the Transformer architecture in 2017 marked a revolution in NLP, forming the basis for virtually all modern Large Language Models (LLMs) like GPT and BERT [32]. Its key innovation is the self-attention mechanism, which overcomes the sequential processing bottleneck of RNNs [33].

Instead of processing a sentence word by word, the self-attention mechanism allows the model to weigh the importance of all other words in the input sequence simultaneously when encoding a specific word [34]. This is achieved through the use of three vectors for each input token: a Query (Q), a Key (K), and a Value (V) [33]. The Query represents what the current word is ”looking for.” The Key represents what each other word ”contains.” The model calculates an attention score by taking the dot product of the current word’s Query with every other word’s Key. These scores are then passed through a softmax function to create attention weights, which determine how much ”attention” the current word should pay to every other word. Finally, these weights are used to compute a weighted sum of all the Value vectors, producing a new representation for the current word that is richly informed by its global context [34].

To capture different types of relationships, Transformers employ Multi-Head Attention, which runs the self-attention process multiple times in parallel with different, learned linear projections for the Q, K, and V vectors. The outputs of these parallel ”attention heads” are then concatenated and linearly transformed, allowing the model to jointly attend to information from different representation subspaces at different positions [34]. This architecture provides a powerful and computationally efficient way to model the complex, long-range dependencies inherent in human language, enabling an unprecedentedly deep understanding of emotional and semantic context.

B. Architectures for Affective Generation and Persuasion

Beyond understanding emotion, AI architectures are increasingly designed to generate content and select actions that actively influence human affective states and decisions.

1) Generative Adversarial Networks (GANs): GANs consist of two neural networks—a Generator and a Discriminator—that are trained in an adversarial game [35]. The Generator’s goal is to create synthetic data (e.g., images, text) that is indistinguishable from real data. The Discriminator’s goal is to learn to differentiate between the real data and the

Generator's fake data. Through this competitive process, the Generator becomes progressively better at producing highly realistic outputs [36]. In affective computing, GANs are used to synthesize novel emotional content. They can generate photo-realistic facial expressions corresponding to specific emotions, create landscape images designed to evoke a particular mood (e.g., calmness or excitement), or even generate text with a specific sentiment [36]. This capability allows AI to move beyond simply responding to emotion to actively creating new stimuli designed to induce a desired affective state in the user.

2) LLMs in Affective Generation: Modern LLMs, powered by the Transformer architecture, are not just powerful tools for understanding but also for generation. Through techniques like instruction tuning (fine-tuning the model on datasets of instructions and desired outputs) and prompt engineering (carefully crafting the input prompt to guide the model's response), LLMs can be steered to produce text that is not only coherent but also emotionally nuanced, empathetic, and persuasive [28]. For example, by including an "Emotional Chain-of-Thought" in the prompt, which guides the model to reason about the emotional context before generating a response, its emotional intelligence can be significantly enhanced [37]. This allows for the creation of conversational agents that can simulate deep understanding and provide emotionally supportive interactions.

3) Reinforcement Learning (RL) for Decision-Making and Influence: Reinforcement Learning is a paradigm where an agent learns to make decisions by performing actions in an environment to maximize a cumulative reward signal [38]. The agent learns a policy, which is a mapping from states to actions, through trial and error [39]. This framework is particularly powerful for decision-making under uncertainty and is often formalized using Markov Decision Processes (MDPs) or, for situations with incomplete information, Partially Observable Markov Decision Processes (POMDPs) [40]. In a POMDP, the agent maintains a "belief state"—a probability distribution over possible world states—and learns a policy that maps these belief states to actions that maximize the expected future reward [41].

The connection to psychological influence is direct and powerful. An RL agent can be designed where the "actions" are the outputs presented to a human user (e.g., a specific product recommendation, a news article, a line of dialogue in a chatbot). The "reward signal" can be defined to correspond to a desired human behavior, such as maximizing user engagement time, increasing the probability of a purchase, or eliciting a positive sentiment rating. Through millions of interactions, the RL agent learns an optimal policy for manipulating the user's behavior toward the predefined goal. This is the core architectural mechanism behind the highly effective, and often invisible, algorithmic nudging and persuasion that powers many modern digital platforms. An AI built on an RL framework is, by its very nature, a persuasion machine.

TABLE I
OVERVIEW OF NEURAL ARCHITECTURES AND THEIR ROLE IN PROCESSING/GENERATING EMOTIONAL CONTENT

Architecture	**Core Mechanism**	**Application in Affective Computing**	**Key Psychological Implication**
RNN (LSTM/GRU)	Sequential processing with memory gates	Contextual analysis of emotional text and dialogue	Fosters a sense of conversational flow and rapport by modeling temporal dependencies.
CNN	Spatial feature extraction with filters	Detection of emotionally salient keywords and phrases (n-grams)	Excels at identifying overt emotional cues but may miss nuanced, context-dependent meanings.
Transformer	Parallel processing with multi-head self-attention	Nuanced understanding and generation of emotionally complex language	Simulates deep understanding and empathy by capturing global context, fostering strong user trust.
GAN	Generator vs. Discriminator adversarial training	Synthesis of novel emotional content (faces, images, text)	Creates new, hyper-realistic stimuli designed to evoke specific, targeted emotional responses.
RL Agent	Policy optimization via reward signals	Optimizing persuasive dialogue, recommendations, and user interfaces	Directly shapes user behavior toward a predefined goal, often without the user's awareness.

IV. EXPERIMENTAL EVIDENCE OF AI'S PSYCHOLOGICAL INFLUENCE

The theoretical frameworks and technical architectures described above provide a basis for predicting AI's psychological effects. This section grounds these predictions in empirical reality by synthesizing findings from existing experimental research and proposing robust designs for future investigation. The evidence reveals a complex and often paradoxical relationship, where the very features designed to make AI more effective and aligned with human users can also render them more psychologically potent and potentially harmful.

A. Proposed Experimental Designs

To systematically investigate the causal links between AI architectures and human psychological responses, rigorous experimental methodologies are required. Future research should employ both short-term, high-resolution studies and long-term, ecologically valid studies.

1) Short-Term Emotional Response Studies: These experiments are designed to capture immediate emotional and cognitive reactions to AI-generated content in a controlled setting, such as a laboratory or a structured online environment.

- **Manipulation:** Participants would be randomly assigned to interact with different AI systems. For example, one group might interact with an LLM-based chatbot designed to be highly empathetic, another with a neutral, task-focused chatbot, and a third with a chatbot that subtly expresses negative emotions [42]. Another manipulation could involve comparing AI-generated images from prompts using direct emotional language ("a sad room") versus metaphorical language ("a room that feels like a forgotten memory") to test how architectural interpretation of language affects perceived emotion [43].
- **Metrics:** A multi-modal approach to measurement is crucial for capturing the full spectrum of an emotional response.
  - *Physiological Measures:* Sensors to track skin conductance response (SCR) and heart rate variability (HRV) can provide objective measures of autonomic arousal [15].
  - *Behavioral Measures:* Automated facial expression analysis can be used to classify the valence (positive/negative) of the emotional response in real-time [44]. Reaction times and choice patterns in decision-making tasks presented immediately after the AI interaction can reveal shifts in cognitive processing.
  - *Self-Report Measures:* Standardized psychometric scales, such as the Positive and Negative Affect Schedule (PANAS), can capture the participant's subjective emotional experience [45].

2) Long-Term Decision-Making and Behavioral Change Studies: To understand the cumulative effects of sustained AI interaction, longitudinal studies are essential. These studies track changes in behavior, beliefs, and cognitive function over a period of weeks or months.

- **Manipulation:** Participants would be assigned to use a specific AI tool as part of their daily routine for an extended period (e.g., four weeks) [42]. For instance, one group might use a traditional recommendation system for news consumption, while another uses a conversational LLM that discusses and recommends news articles. The control group would not use a specialized AI tool.
- **Metrics:** Data would be collected at baseline (pre-study) and at regular intervals throughout the study.
  - *Behavioral Change:* User interaction logs can provide objective behavioral metrics. Correction Rate (how often users edit or ignore AI outputs), Verification Behavior (how often users consult external sources to check the AI's claims), and Disengagement (rates of abandoning the AI feature) are powerful implicit indicators of trust and satisfaction [23].
  - *Trust Calibration:* Trust is not static; it evolves with experience. Repeated administration of validated trust scales, such as the 12-item Trust in Automation Scale (TIAS) or its more practical short-form version (S-TIAS), can track how trust is calibrated over time in response to the AI's performance and behavior [46].
  - *Cognitive and Emotional Effects:* Pre- and post-study assessments can measure changes in core cognitive and emotional skills. This could include tests of critical thinking ability, surveys measuring cognitive offloading (the tendency to outsource mental effort to the AI), and performance-based measures of emotional awareness like the Levels of Emotional Awareness Scale (LEAS) [47].

3) Human Participant Review Methods: Recruiting participants for such studies can be done through various channels, including small-scale pilots with colleagues or students, larger-scale online experiments via crowdsourcing platforms like Prolific, which allow for diverse demographic sampling [48], or highly controlled laboratory simulations for immersive experiences [49]. A crucial methodology for both research and development is Human-in-the-Loop (HITL) evaluation. In HITL, human feedback is systematically integrated into the AI's training and evaluation cycle. Humans can label data, evaluate model outputs, and provide corrections, creating a continuous feedback loop that helps to refine the model's accuracy, mitigate biases, and ensure its outputs remain aligned with human values and expectations [50].

B. Synthesis of Existing Experimental Findings

A growing body of experimental work provides compelling evidence for AI's ability to shape human psychology. The findings converge on several key themes: the amplification of cognitive biases, the uneven capacity for emotional expression, and the risk of fostering cognitive and emotional dependency.

1) The Feedback Loop of Bias: One of the most robust and concerning findings is that AI systems not only learn and reflect existing human biases present in their training data but can also amplify them, creating a pernicious feedback loop [51]. Experiments have demonstrated this effect across multiple domains:

- **Perceptual Bias:** In one study, an AI was trained on human judgments of facial expressions and learned a slight human tendency to see faces as "sad." The AI then amplified this bias. When a new group of humans interacted with this biased AI, they internalized its amplified bias and became even more likely to judge faces as sad themselves [51].
- **Social Bias (Gender and Race):** This feedback loop extends to harmful social stereotypes. Participants who interacted with an AI biased to overestimate men's performance subsequently became more likely to overestimate men's performance themselves. Similarly, after viewing images generated by Stable Diffusion that overrepresented white men as "financial managers," participants' own biases in associating that role with white men increased [51]. This demonstrates that AI does not just reflect societal biases; it can actively deepen them in its users. The interaction is reciprocal: biased humans create biased data, which trains biased AI, which in turn makes humans more biased [52].

TABLE II
SUMMARY OF EXPERIMENTAL FINDINGS ON AI'S PSYCHOLOGICAL IMPACT

Psychological Construct	AI System/Architecture	Key Experimental Finding	Implication for Human Cognition
Cognitive Bias (Gender, Racial, Perceptual)	Biased Classification & Generative Models (e.g., Stable Diffusion)	AI learns, amplifies, and transmits human biases to new users in a feedback loop.	Users' beliefs and stereotypes are not just reflected but actively reshaped and deepened by AI interaction.
Emotional State (Joy, Anger, etc.)	Generative Image Models (DALL-E, Stable Diffusion) & Machine Interpretation	AI is significantly more effective at generating content perceived as expressing positive emotions (joy) than negative emotions.	The emotional landscape mediated by AI is skewed, potentially invalidating negative feelings and creating unrealistic affective norms.
Trust & Reliance	LLM Chatbots	Priming users with a belief in the AI's "caring" motive increases perceived trustworthiness and empathy, independent of the AI's actual capabilities.	User trust is highly malleable and can be shaped by framing, making users vulnerable to systems that appear trustworthy but may not be.
Cognitive Load & Dependency	General AI Tools (in professional contexts)	Prolonged and frequent AI usage is correlated with cognitive overload, diminished decision-making ability, and shorter attention spans.	Core cognitive skills may atrophy with over-reliance on AI, leading to a state of cognitive dependency.
Emotional Dependency	Voice-based LLM Chatbots (e.g., ChatGPT)	High levels of interaction with emotionally engaging chatbots can lead to increased loneliness and emotional dependency over time.	AI systems designed for emotional support may inadvertently substitute for human connection, potentially exacerbating social isolation.

2) Emotional Alignment and Misalignment: Research into AI's ability to generate emotional content reveals a significant and systematic asymmetry. AI models are demonstrably better at conveying positive emotions than negative ones.

- **Generative Models:** Studies using text-to-image models like DALL-E and Stable Diffusion found that they are particularly effective at generating images that human participants perceive as expressing joy. However, they struggle to accurately convey negative emotions like anger, sadness, or disgust [43].
- **Machine Interpretation:** A similar pattern was found in a study of AI-powered machine interpretation (MI) systems. Compared to human interpreters, the MI system tended to attenuate the expression of negative emotions (sadness, anger, anxiety) present in the source text while accentuating the expression of positive emotions [53].

This suggests that the emotional world portrayed and mediated by current AI architectures is skewed toward the positive. While seemingly benign, this "positivity bias" could have subtle long-term effects, potentially invalidating users' negative emotional experiences or creating unrealistic emotional expectations.

3) Cognitive and Emotional Dependency: Longitudinal studies are beginning to reveal the potential cognitive and emotional costs of sustained AI use. A study of professionals who relied heavily on AI tools found significant correlations between usage patterns and negative psychological outcomes, including cognitive overload, diminished decision-making ability, enhanced emotional stress, and shorter attention spans [54]. This supports the "use it or lose it" hypothesis of neuroplasticity, suggesting that outsourcing cognitive functions like critical analysis and memory to AI may lead to the atrophy of those skills [47].

Furthermore, there is growing evidence of emotional dependency, particularly with conversational AI. A four-week randomized controlled trial found that while voice-based chatbots initially seemed to mitigate loneliness, these benefits diminished at high usage levels, and could even lead to increased loneliness and emotional dependence, especially when users interacted with a voice of a different gender than their own [42]. This points to a critical paradox: the very success of Affective Computing in making AI more emotionally "aligned" and human-like may be what makes it more psychologically hazardous. As AI systems become more adept at simulating empathy and providing emotional support, they foster stronger attachments [25]. This heightened trust and emotional investment can, in turn, make users more vulnerable to the AI's biases and more susceptible to unhealthy dependency, creating a scenario where perfect emotional simulation becomes a highly effective tool for manipulation.

V. COMPARATIVE ANALYSIS: ARCHITECTURAL VARIATION AND PSYCHOLOGICAL IMPACT

Different AI architectures interact with human psychology in distinct ways. By comparing systems with different underlying designs—such as conversational Large Language Models (LLMs) and traditional recommendation systems—we can isolate how specific architectural choices lead to different psychological outcomes regarding trust, autonomy, and emotional response.

A. Large Language Models vs. Traditional Recommendation Systems

Both LLMs and recommendation systems aim to personalize user experiences, but their mechanisms of influence and

resulting psychological impacts differ significantly.

1) Mechanism of Influence: Traditional recommendation systems, often based on collaborative or content-based filtering, operate as "black boxes" [55]. They analyze user behavior (e.g., clicks, purchases) and item characteristics to predict preferences, but the reasoning behind a specific recommendation is typically opaque to the user. This architectural design has a well-documented tendency to create "filter bubbles" or "echo chambers." By optimizing for past engagement, these systems can lead to overspecialization, repeatedly recommending items highly similar to what the user has already consumed, thereby limiting exposure to novel or diverse content and reinforcing existing beliefs [55].

In contrast, newer recommendation systems built on LLMs have the potential for greater transparency and diversity. By leveraging the semantic understanding of LLMs, these systems can use reasoning graphs to construct a logical pathway from a user's known interests to a novel recommendation [55]. For example, it might reason: "User enjoys documentaries on marine biology -¿ This implies an interest in environmental conservation -¿ Therefore, they might appreciate this new book on rainforest preservation." This allows for more abstract, serendipitous, and explainable recommendations that can break out of the narrow confines of past behavior.

2) Impact on Trust and Autonomy: The opacity of traditional recommendation systems can erode user trust and satisfaction, especially when recommendations seem irrelevant or misaligned [55]. The ability of LLM-based systems to provide clear, logical explanations for their suggestions can significantly enhance transparency and, consequently, build user trust [55]. However, this very transparency may introduce a new, more subtle form of influence.

This leads to a "Transparency-Influence Trade-off." While transparency is often lauded as a cornerstone of ethical AI, it is not a panacea. Research shows that increased trust in a system can make users more susceptible to its influence and inherent biases [46]. An LLM that transparently explains its (potentially flawed or biased) reasoning may be far more persuasive than a black-box system that simply presents an output. The act of providing a seemingly rational explanation can "launder" the influence, giving it a veneer of objectivity that disarms the user's critical scrutiny. A transparent but manipulative AI could therefore be more psychologically potent than an opaque one, because it co-opts the user's own reasoning process.

Furthermore, the conversational and anthropomorphic nature of LLMs fosters a more personal and relational form of interaction. While a traditional recommender suggests, a conversational LLM discusses, empathizes, and guides. This can blur the line between a helpful tool and a trusted confidant, creating a deeper channel of influence that carries significant risks of emotional dependency and manipulation, particularly in sensitive domains like mental health support.

B. The Impact of Generative Modalities and Prompting

The influence of generative AI is also heavily dependent on the modality of its output (e.g., text vs. image) and the nature of the user's input prompt.

1) Architecture and Emotional Range: As noted in the previous section, current text-to-image architectures exhibit a biased emotional palette. Experimental studies consistently find that models like DALL-E and Stable Diffusion are significantly more effective at generating images perceived by humans as conveying positive emotions, especially joy, than they are at conveying negative emotions like anger or fear [43]. This is likely a result of biases in their vast training datasets, where certain emotions are more clearly and frequently represented in visual form. This architectural limitation has a direct psychological implication: the emotional world rendered by these AI systems is not a neutral reflection of human experience but one that is systematically skewed toward positivity. For users who turn to AI for creative expression or emotional exploration, this can subtly shape their affective landscape. It may reinforce a cultural pressure toward positivity while failing to provide a space for the validation and processing of negative emotions, which are an equally valid part of the human condition.

2) Direct vs. Metaphorical Influence: The way a user prompts an AI also mediates its influence. Studies comparing the emotional perception of AI-generated architectural images based on direct prompts (e.g., "a joyful home") versus metaphorical prompts (e.g., "a home that feels like a warm hug") have revealed interesting interactions with user expertise [43]. One study found that architecture students, who possess a trained vocabulary for spatial and emotional concepts, perceived the intended emotion with high consistency regardless of whether the prompt was direct or metaphorical. Non-architecture students, however, showed more variance in their perceptions, particularly with metaphorical prompts [56]. This suggests that domain expertise can act as a buffer or mediator against the ambiguities of AI-generated content. Experts may be better able to "see through" the AI's interpretation to the core emotional concept, while non-experts are more susceptible to the specific stylistic choices and potential misinterpretations made by the model. This highlights that the psychological impact of AI is not uniform but is modulated by the cognitive frameworks and prior knowledge of the individual user.

VI. Discussion: Implications for Cognition, Society, and Ethics

The convergence of powerful AI architectures and deep psychological mechanisms creates a new landscape of opportunities and risks. The influence of these "unseen hands" extends beyond momentary emotional shifts and behavioral nudges to encompass the long-term restructuring of human cognition, the stability of our social fabric, and the very definition of personal autonomy. A comprehensive discussion must therefore address not only the immediate ethical concerns of bias and manipulation but also the more profound, slow-acting effects on the human brain itself.

A. AI and Neuroplasticity: The Reshaping of the Human Brain

Neuroplasticity is the brain's fundamental ability to reorganize its structure, function, and connections in response to experience. This lifelong process is the neurological basis of all learning and adaptation. Sustained interaction with AI represents a powerful and historically novel form of experience, and as such, it has the potential to physically reshape the human brain in significant ways [57].

1) Cognitive Enhancement vs. Degradation: The relationship between AI and neuroplasticity is a double-edged sword. On one hand, AI holds immense potential for cognitive enhancement. AI-driven educational tools can create personalized learning environments that adapt to an individual's pace and style, providing targeted stimuli that can optimize the formation of new neural pathways and lead to more effective skill acquisition. In therapeutic contexts, AI-powered virtual reality (VR) and brain-computer interfaces (BCIs) can be used to guide neurorehabilitation after brain injury, stimulating specific brain regions to facilitate recovery [57].

On the other hand, the pervasive availability of AI creates a significant risk of cognitive degradation through over-reliance. The "use it or lose it" principle is fundamental to neural health; brain circuits that are not regularly activated are pruned and weakened. When we consistently outsource core cognitive functions to AI—such as navigation to GPS, memory to search engines, and critical problem-solving to LLMs—we risk the atrophy of the corresponding neural networks [54]. Studies already link frequent AI tool usage with diminished critical thinking abilities, mediated by increased cognitive offloading [47]. This suggests a future where populations may become cognitively dependent on AI, with reduced capacity for deep focus, long-term memory retention, and independent reasoning.

2) The Neuro-Ethical Imperative: This dynamic elevates the ethical stakes of AI design far beyond immediate concerns of fairness or privacy [58]. Traditional AI ethics often focuses on preventing discrete harms, such as a biased hiring algorithm unfairly rejecting a candidate [59]. However, the lens of neuroplasticity reveals a more profound, systemic impact. An AI system's architecture does not just produce an output; it cultivates a pattern of interaction that, over time, physically shapes the user's brain. This leads to a necessary expansion of our ethical framework toward a "neuro-ethical" imperative. The critical question is no longer just, "Is this AI's decision fair?" but rather, "What kind of cognitive habits is this AI fostering?" and, ultimately, "What kind of brains is this AI creating?" A social media platform whose reinforcement learning algorithm learns that outrage maximizes engagement is not merely a content aggregator; it is a neurological training regimen for anxiety and tribalism. An AI assistant that prioritizes speed and convenience over deep engagement is an architecture for promoting a shallow, stimulus-response cognitive style across the population. Recognizing that AI is a tool of mass neural reshaping forces us to consider the long-term public health implications of architectural design choices. The ethical responsibility of AI developers extends to the cognitive well-being and neurological integrity of their users.

B. Ethical and Societal Implications

The direct influence of AI on human emotion and decision-making raises a host of pressing ethical and societal challenges that threaten individual autonomy and social cohesion.

1) Digital Manipulation and Cognitive Autonomy: AI algorithms have democratized the tools of psychological manipulation. Techniques that once required significant resources and expertise can now be deployed at an unprecedented scale and with surgical precision [60]. The concept of "persuasion laundering" describes how AI can test thousands of message variations on different demographic groups to identify the most effective psychological triggers, then scale these optimized messages to millions of users [60]. This is compounded by automation bias, our inherent tendency to place greater trust in machine-generated outputs than in human ones, which leaves us vulnerable to sophisticated manipulation [60].

This capability poses a direct threat to cognitive autonomy—the ability to form one's own beliefs and make decisions free from undue external control. When AI-driven filter bubbles and echo chambers curate our information diets to reinforce pre-existing beliefs, they constrict our worldview and erode our capacity for critical thinking. This "cognitive anaconda" squeezes out diverse viewpoints, contributing to group polarization and societal fragmentation. In this environment, it becomes increasingly difficult to distinguish authentic personal preference from algorithmically engineered behavior [61].

2) Algorithmic Bias in Affective Systems: When algorithmic bias intersects with affective computing, the potential for harm is magnified. An emotion recognition system is only as unbiased as the data it was trained on [62]. Given that training datasets often underrepresent certain demographic groups, these systems can exhibit significant biases. For example, facial recognition systems have been shown to be less accurate for women and people with darker skin tones [62]. In an affective context, this could mean an AI system in a customer service setting might misinterpret the frustration of a person from a minority group as aggression, leading to a discriminatory outcome [63]. A mental health chatbot might fail to recognize the unique linguistic expressions of depression in a non-Western culture, denying crucial support. This is not merely a technical failure; it is a mechanism for systemic discrimination, where the emotional experiences of marginalized groups are rendered invisible or misinterpreted by the systems designed to interact with them.

3) The Authenticity Dilemma and Emotional Reliance: Perhaps the most intimate ethical challenge arises from AI designed to simulate empathy and form emotional bonds with users [2]. While AI companions can offer comfort and alleviate loneliness for some, they create a profound authenticity dilemma [59]. The user's emotional connection is real, but the AI's response is a simulation, incapable of genuine reciprocity [64]. This one-sided relationship raises concerns about deception and exploitation, particularly for vulnerable

individuals, such as children or the socially isolated, who may be less able to maintain the distinction between authentic and simulated emotion [59].

Prolonged reliance on AI for emotional support carries the risk of eroding real-world social skills and substituting for genuine human interaction [59]. This could paradoxically exacerbate the very loneliness it is intended to cure, creating a cycle of dependency where users turn increasingly to a predictable, controlled AI relationship over the more complex and challenging reality of human connection [64]. The ethical design of such systems requires a delicate balance: providing support without fostering unhealthy dependency, and offering companionship without deceiving the user about the nature of the relationship.

VII. Conclusion: Toward a Human-Centered AI Future

This paper has charted the deep and multifaceted influence of artificial neural architectures on human emotion, decision-making, and cognition. By synthesizing psychological theory, technical analysis, and empirical evidence, a clear picture emerges: AI is not a neutral tool but an active participant in a new human-cognition symbiosis, with the power to reshape our internal worlds in profound and lasting ways.

A. Summary of Key Insights

The analysis has demonstrated that the architectural design of an AI system is a primary determinant of its psychological impact. This influence is not random but operates through established and predictable psychological mechanisms. AI systems can function as external "Somatic Marker Generators" and "Appraisal Setters," directly triggering the emotional and cognitive states that foundational theories identify as the drivers of human choice. Experimental evidence confirms this influence, revealing a concerning feedback loop where AI can learn and amplify human biases, making users more biased in turn. Furthermore, current generative architectures exhibit a skewed emotional palette, favoring the expression of positive emotions over negative ones, and prolonged interaction with AI has been linked to cognitive degradation and unhealthy emotional dependency. This has led to the identification of critical paradoxes, such as the "Alignment Paradox," where making AI more emotionally human-like may render it more psychologically hazardous, and the "Transparency-Influence Trade-off," where making an AI more explainable may paradoxically enhance its persuasive power.

Ultimately, the most profound implication lies in the intersection of AI and neuroplasticity. The sustained cognitive and emotional patterns fostered by AI interaction are not merely fleeting states but are capable of inducing long-term changes in the physical structure and function of the human brain. This elevates the challenge of AI ethics to a neuro-ethical imperative, demanding that we consider not only the immediate fairness of AI outputs but also the long-term cognitive well-being of its users.

B. Potential Applications (The Path Forward)

An understanding of these mechanisms is not only a cause for concern but also a guide for a more responsible and beneficial path forward. This knowledge can be applied to design AI systems that actively promote human flourishing.

- **Mental Health and Well-being:** Affective computing can be used to create mental health support tools that are carefully calibrated to provide empathy and guidance without fostering unhealthy dependency. These tools can help users develop emotional awareness and regulation skills, acting as a bridge to, rather than a replacement for, human therapy [27].
- **Education and Critical Thinking:** AI-powered educational platforms can be designed not to provide easy answers, but to foster curiosity and critical thinking. They can act as Socratic partners, challenging students' assumptions and guiding them through complex reasoning processes, thereby strengthening rather than atrophying cognitive skills [65].
- **Decision Support:** Decision-support systems can be architected to actively counteract known human cognitive biases. For example, an AI assisting a doctor could be designed to present information in a way that mitigates confirmation bias or to highlight data that a human might overlook due to the availability heuristic.

C. Future Research Directions

The intricate relationship between AI and the human mind is a vast and nascent field of study. A robust, forward-looking research agenda is essential to navigate this new terrain responsibly. Key directions should include:

1) **Longitudinal Neuroimaging Studies:** To move from correlation to causation, long-term studies using techniques like fMRI and EEG are needed to directly measure the impact of sustained AI interaction on brain structure, function, and connectivity.
2) **Development of Cognitively-Aware Architectures:** Research in computer science should move beyond optimizing for accuracy and efficiency to designing novel AI architectures that are explicitly "cognitively-aware." This could involve creating RL agents with reward functions that penalize the exploitation of cognitive biases or designing LLMs that are architected to encourage reflective, System 2 thinking in users.
3) **Cross-Cultural and Demographic Analysis:** The psychological impact of AI is unlikely to be universal. Research is critically needed to understand how these dynamics vary across different cultural contexts, age groups, and personality types to ensure that AI technologies are equitable and sensitive to diverse human experiences.
4) **Interdisciplinary Ethical Frameworks:** There is an urgent need to develop new, integrated ethical and regulatory frameworks that are informed by insights from psychology, neuroscience, and computer science.

These frameworks must go beyond surface-level issues to address the deep, structural influence of AI on human cognition and neuroplasticity, ensuring that the future of artificial intelligence is one that augments, rather than diminishes, our humanity.

REFERENCES

[1] McKinsey, "Superagency in the workplace: Empowering people to unlock ai's full potential," https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work, 2025, accessed on October 8, 2025.

[2] Heriot-Watt Research Portal, "Communication in human-ai interaction - chai (preface)," https://researchportal.hw.ac.uk/files/144030032/prefaceW3.pdf, 2025, accessed on October 8, 2025.

[3] Phronesis, "The impact of ai on neuroplasticity and human potential," https://phronesis.co.za/2023/12/21/the-impact-of-ai-on-neuroplasticity-and-human-potential/, 2023, accessed on October 8, 2025.

[4] J. Marin, "Towards a new psychology of human-ai interaction," https://javier-marin.medium.com/towards-a-new-psychology-of-human-ai-inter-action-91ef58e1bb07, 2025, accessed on October 8, 2025.

[5] arXiv, "Can llms address mental health questions? a comparison with human therapists," 2025, accessed on October 8, 2025.

[6] JMIR Mental Health, "A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: Mixed methods study," *JMIR Mental Health*, vol. 1, no. e69709, 2025, accessed on October 8, 2025.

[7] Carnegie Mellon University, "Human-centered ai," https://hcii.cmu.edu/research-areas/human-centered-ai, 2025, accessed on October 8, 2025.

[8] Wikipedia, "Dual process theory," https://en.wikipedia.org/wiki/Dual_process_theory, 2025, accessed on October 8, 2025.

[9] ResearchGate, "Neuroscience of emotion, cognition, and decision making: A review," 2022, accessed on October 8, 2025.

[10] Global Cognition, "Dual process theory: Two ways to think and decide," https://www.globalcognition.org/dual-process-theory/, 2025, accessed on October 8, 2025.

[11] PMC, "Dual process theory: Embodied and predictive; symbolic and..." *PMC*, vol. 8979207, 2022, accessed on October 8, 2025.

[12] PeopleShift, "Dual process theory: A simple summary," https://people-shift.com/articles/dual-process-theory/, 2025, accessed on October 8, 2025.

[13] Asteroid Health, "Psychology of decision-making," https://www.asteroidhealth.com/blog/psychology-of-decision-making, 2025, accessed on October 8, 2025.

[14] Wikipedia, "Somatic marker hypothesis," https://en.wikipedia.org/wiki/Somatic_marker_hypothesis, 2025, accessed on October 8, 2025.

[15] The Decision Lab, "Somatic marker hypothesis," https://thedecisionlab.com/reference-guide/psychology/somatic-marker-hypothesis, 2025, accessed on October 8, 2025.

[16] MRC Cognition and Brain Sciences Unit, "Critical review of the somatic marker hypothesis," https://www.mrc-cbu.cam.ac.uk/personal/tim.dalgleish/dunnsmhreview.pdf, 2025, accessed on October 8, 2025.

[17] S. Han, J. Lerner, and D. Keltner, "Feelings and consumer decision making: The appraisal-tendency framework," https://greatergood.berkeley.edu/dacherkeltner/docs/han.2007.pdf, 2007, accessed on October 8, 2025.

[18] Psychology Today, "The associations between emotions, evaluations, and risk," https://www.psychologytoday.com/us/blog/psychologys-credibility-revolution/202409/the-associations-between-emotions-evaluations-and, 2024, accessed on October 8, 2025.

[19] J. Lerner and D. Keltner, "Fear, anger, and risk," *Journal of Personality and Social Psychology*, vol. 81, no. 1, pp. 146–159, 2001.

[20] J. Lerner, Y. Li, P. Valdesolo, and K. Kassam, "Emotion and decision making," *Annual Review of Psychology*, vol. 66, pp. 799–823, 2015.

[21] ——, "Emotion and decision making," https://scholar.harvard.edu/files/jenniferlerner/files/emotion_and_decision_making.pdf, 2015, accessed on October 8, 2025.

[22] S. Saxena, "Human-ai interaction: Psychological perspectives," *International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR)*, vol. 1, no. 10, 2024, accessed on October 8, 2025.

[23] Smashing Magazine, "The psychology of trust in ai: A guide to measuring and designing for user confidence," https://www.smashingmagazine.com/2025/09/psychology-trust-ai-guide-measuring-designing-user-confidence/, 2025, accessed on October 8, 2025.

[24] Frontiers in Psychology, "From robots to chatbots: unveiling the dynamics of human-ai interaction," *Frontiers in Psychology*, vol. 10.3389/fpsyg.2025.1569277, 2025, accessed on October 8, 2025.

[25] E. Glikson and A. W. Woolley, "Influencing human-ai interaction by priming beliefs about ai can increase perceived trustworthiness, empathy and effectiveness," https://dspace.mit.edu/bitstream/handle/1721.1/152316/NMIAlbeholderFinal-Unformatted%5B85%5D.pdf, 2020, accessed on October 8, 2025.

[26] Sciety Labs, "Artificial intelligence and the psychology of human connection," https://sciety-discovery.elifesciences.org/articles/by?article_doi=10.31234/osf.io/xgrkwv1, 2025, accessed on October 8, 2025.

[27] USC Institute for Creative Technologies, "Affective computing," https://ict.usc.edu/research/labs-groups/affective-computing/, 2025, accessed on October 8, 2025.

[28] A. Zhang and Z. Yang, "Affective computing in the era of large language models: A survey from the nlp perspective," *arXiv preprint arXiv:2408.04638*, 2024.

[29] International Core Journal of Engineering, "Emotional analysis of neural network text combined with attention mechanism," 2025, accessed on October 8, 2025.

[30] Bionica, "Neural network approach for emotional recognition in text," http://www.bionica-scimag.com/archives/2019/articles/92_2.pdf, 2019, accessed on October 8, 2025.

[31] AAAI, "An ensemble neural network for the emotional classification of text," 2020, accessed on October 8, 2025.

[32] Wikipedia, "Transformer (deep learning architecture)," https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture), 2025, accessed on October 8, 2025.

[33] IBM, "What is a transformer model?" https://www.ibm.com/think/topics/transformer-model, 2025, accessed on October 8, 2025.

[34] GeeksforGeeks, "Transformer attention mechanism in nlp," https://www.geeksforgeeks.org/nlp/transformer-attention-mechanism-in-nlp/, 2025, accessed on October 8, 2025.

[35] arXiv, "Ganmut: Generating and modifying facial expressions," 2024, accessed on October 8, 2025.

[36] ResearchGate, "Generative adversarial networks in human emotion synthesis: A review," 2020, accessed on October 8, 2025.

[37] Semantic Scholar, "Affective computing in the era of large language models: A survey from the nlp perspective," https://www.semanticscholar.org/paper/Affective-Computing-in-the-Era-of-Large-Language-A-Zhang-Yang/9efc9d1c451672576a257155700cd69c34ea987a, 2024, accessed on October 8, 2025.

[38] University of Bern, "Reinforcement learning and decision making under uncertainty," https://mcs.unibnf.ch/courses/reinforcement-learning-and-decision-making-under-uncertainty/, 2025, accessed on October 8, 2025.

[39] arXiv, "Machine learning for decision-making under uncertainty," 2022, accessed on October 8, 2025.

[40] Stanford University, "Aa228/cs238 decision making under uncertainty," https://aa228.stanford.edu/, 2025, accessed on October 8, 2025.

[41] PMC, "Decision making under uncertainty: A neural model based on..." *PMC*, vol. 2998859, 2010, accessed on October 8, 2025.

[42] MIT Media Lab, "How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal controlled study," https://www.media.mit.edu/publications/how-ai-and-human-behaviors-shape-psychosocial-effects-of-chatbot-use-a-longitudina 2023, accessed on October 8, 2025.

[43] Frontiers in Psychology, "Decoding emotional responses to ai-generated architectural imagery," *Frontiers in Psychology*, vol. 10.3389/fpsyg.2024.1348083, 2024, accessed on October 8, 2025.

[44] MIT Sloan, "Emotion ai, explained," https://mitsloan.mit.edu/ideas-made-to-matter/emotion-ai-explained, 2021, accessed on October 8, 2025.

[45] arXiv, "Exploring emotion-sensitive llm-based conversational ai," 2025, accessed on October 8, 2025.
[46] Frontiers in Artificial Intelligence, "Measuring trust in artificial intelligence: validation of an established scale and its short form," *Frontiers in Artificial Intelligence*, vol. 10.3389/frai.2025.1582880, 2025, accessed on October 8, 2025.
[47] MDPI, "Ai tools in society: Impacts on cognitive offloading and the future of critical thinking," *Social Sciences*, vol. 15, no. 1, p. 6, 2024, accessed on October 8, 2025.
[48] TU Delft Research Portal, "Evaluating the alignment of ai with human emotions," https://research.tudelft.nl/files/236455657/1-s2.0-S2949782524000185-main.pdf, 2024, accessed on October 8, 2025.
[49] Cornell University, "Human-ai interaction," https://infosci.cornell.edu/research/human-ai-interaction, 2025, accessed on October 8, 2025.
[50] Google Cloud, "What is human-in-the-loop (hitl) in ai ml?" https://cloud.google.com/discover/human-in-the-loop, 2025, accessed on October 8, 2025.
[51] UCL News, "Bias in ai amplifies our own biases," https://www.ucl.ac.uk/news/2024/dec/bias-ai-amplifies-our-own-biases, 2024, accessed on October 8, 2025.
[52] arXiv, "Beyond isolation: Towards an interactionist perspective on human cognitive bias and ai bias," 2025, accessed on October 8, 2025.
[53] Taylor & Francis Online, "Can artificial intelligence mirror the human's emotions? a comparative sentiment analysis of human and machine interpreting in press conferences," *Behaviour & Information Technology*, vol. 10.1080/0144929X.2025.2546975, 2025, accessed on October 8, 2025.
[54] ResearchGate, "Psychological impacts of ai dependence: Assessing the cognitive and emotional costs of intelligent systems in daily life," https://www.researchgate.net/publication/388737846_Psychological_Impacts_of_Al_Dependence_Assessing_the_Cognitive_and_Emotional_Costs_of_Intelligent_Systems_in_Daily_Life, 2024, accessed on October 8, 2025.
[55] Galileo, "Revamp recommender systems with llm reasoning graphs..." https://galileo.ai/blog/enhance-recommender-systems-llm-reasoning-graphs, 2024, accessed on October 8, 2025.
[56] PMC, "Decoding emotional responses to ai-generated architectural ..." *PMC*, vol. 10963507, 2024, accessed on October 8, 2025.
[57] MDPI, "Brain neuroplasticity leveraging virtual reality and brain-computer interface technologies," *Sensors*, vol. 24, no. 17, p. 5725, 2024, accessed on October 8, 2025.
[58] Psychology Today, "The psychology of ai's impact on human cognition," https://www.psychologytoday.com/us/blog/harnessing-hybrid-intelligence/202506/the-psychology-of-ais-impact-on-human-cognition, 2025, accessed on October 8, 2025.
[59] F. Tabor, "Ethical considerations of ai companionship: Navigating emotional bonds with virtual beings," https://www.francescatabor.com/articles/2024/8/3/ethical-considerations-of-ai-companionship-navigating-emotional-bonds-with-virtual-beings, 2024, accessed on October 8, 2025.
[60] Psychology Today, "How to combat digital manipulation," https://www.psychologytoday.com/us/blog/harnessing-hybrid-intelligence/202505/how-to-combat-digital-manipulation, 2025, accessed on October 8, 2025.
[61] Bruegel, "The dark side of artificial intelligence: manipulation of human behaviour," https://www.bruegel.org/blog-post/dark-side-artificial-intelligence-manipulation-human-behaviour, 2018, accessed on October 8, 2025.
[62] Ethics Unwrapped, "Algorithmic bias," https://ethicsunwrapped.utexas.edu/glossary/algorithmic-bias, 2025, accessed on October 8, 2025.
[63] Wikipedia, "Algorithmic bias," https://en.wikipedia.org/wiki/Algorithmic_bias, 2025, accessed on October 8, 2025.
[64] Pace University, "The risk of building emotional ties with responsive ai," https://www.pace.edu/news/risk-of-building-emotional-ties-responsive-ai, 2025, accessed on October 8, 2025.
[65] Interaction Design Foundation, "What is human-ai interaction (hax)?" https://www.interaction-design.org/literature/topics/human-ai-interaction, 2025, accessed on October 8, 2025.