

DOI: https://doi.org/10.48009/2_iis_2026_148

Bridging the affective-informational gap: A temporal agentic framework for enhancing human cognitive profiles

Gianluca Zanella, *Augusta University, School of Computer & Cyber Sciences*, gzanella@augusta.edu
Pouria Rad, *Augusta University, School of Computer & Cyber Sciences*, peslamirad@augusta.edu

Abstract

As Generative Artificial Intelligence (GenAI) systems transition to autonomous agents, establishing calibrated human trust is critical. However, under the Computers are Social Actors (CASA) paradigm, highly conversational interfaces trigger the Affective-Informational Paradox: users mindlessly anthropomorphize AI, leading to trust miscalibration and cognitive offloading. Concurrently, baseline predictive models struggle to account for the bounded rationality and noise inherent in real-world human choice. To address these systemic vulnerabilities, this paper introduces the Temporal Agentic Framework, a socio-technical architecture grounded in the Tensions-Based View (TBV). We propose a rigorous three-agent pipeline that leverages Cognitive Model Priors to internalize human heuristic noise, demonstrating a 49.5% reduction in predictive error (Brier Score) and achieving an AUC-ROC of 0.895. Governed by an Offline Reinforcement Learning (ORL) Temporal Critic, the system dynamically rewrites its temporal framing to match a user's time orientation profile. This Adaptive Affective Fit deliberately modulates emotional tension, a design goal aimed at restoring human epistemic vigilance. In proof-of-concept evaluation, the agentic feedback loop increased the framework's internally validated convergence rate to 94.0% and improved the estimated causal behavioral shift (ATE) by 92.9%, outcomes that await validation with human participants. Ultimately, this research contributes a design science Information Systems (IS) blueprint, supported by proof-of-concept evidence, for designing trustworthy AI intended to safeguard cognitive autonomy in high-stakes environments.

Keywords: Computers are Social Actors, Tensions-Based View, Design Science Research, Affective-Informational Paradox, Agentic AI, Cognitive Model Priors, Temporal Framing

Introduction

The rapid integration of Generative Artificial Intelligence (AI) into daily applications has shifted the paradigm of human-computer interaction from mere tool usage to collaborative partnerships (McWilliams et al., 2025; Romeo & Conti, 2026). As organizations increasingly deploy Agentic AI systems, capable of independent reasoning, planning, and execution, the fundamental bottleneck to successful adoption is no longer technical capability, but a matter of human trust (Jiang et al., 2025). Establishing appropriate trust is essential, however, the deployment of highly conversational, Multi-Modal AI introduces a critical vulnerability known as the Affective-Informational Paradox. When users instinctively anthropomorphize engaging AI interfaces, it frequently leads to pseudo-intimacy, trust miscalibration, diminished epistemic vigilance, and over-reliance through cognitive offloading (Placani, 2024), which could cause severe consequences.

Whereas automation bias describes a momentary tendency to over-rely on system output within a single decision (Romeo & Conti, 2026), the Cognitive Atrophy Paradox names its cumulative, longitudinal counterpart: the habitual offloading of analytical effort that gradually degrades the underlying capacity itself (Kabashkin, 2025; Mahajan, 2025). It is therefore best understood as a dispositional and developmental construct rather than a point-in-time reliance error. Furthermore, this over-reliance frequently manifests in clinical and high-stakes settings as false confirmation errors, wherein humans uncritically accept an AI's erroneous agreement with their initial biases, bypassing analytical reflection even when Explainable AI (XAI) is provided (Rosenbacke, 2024, 2025). If Agentic AI is to be safely utilized in complex domains such as cybersecurity, healthcare, and teamwork management, systems must be designed to align with the intricate realities of human cognitive processing and mitigate automation bias (Romeo & Conti, 2026; Yu et al., 2017). Fundamentally, systems must augment human judgment while minimizing the introduction of cognitive biases and mitigating deviations from bounded rationality.

Indeed, the study of human decision-making under uncertainty has revealed that individuals rarely act as perfectly rational utility maximizers (Von Neumann & Morgenstern, 1947). Foundational theories demonstrate that human choices are systematically skewed by cognitive biases such as pessimism, regret minimization, and the disproportionate weighting of rare events (Erev et al., 2017; Kahneman & Tversky, 2013). Furthermore, recent research highlights a profound dataset bias in machine-learned theories of economic decision-making, with a study finding that human behavior in controlled laboratory settings differs significantly from decisions made in noisy, real-world or crowdsourced environments (Thomas et al., 2024). This is particularly true when individuals interact with online social media or AI chatbots, where boundaries are blurred and cognitive biases are often exposed and exploited. For example, users tend to overestimate the benefits of online disclosure of sensitive information (instant gratification) while discounting the long-term risks of privacy/security breaches (Hallam & Zanella, 2017).

Indeed, time perspective is a central dimension of human cognitive variability because it dictates how individuals weigh short-term outcomes against long-term impacts, which affects also education and training (Orbell et al., 2004). When AI agents present information or recommend actions, the temporal framing of that information, whether emphasizing immediate relief or long-term growth, dramatically alters the user's perception of utility and, consequently, their trust in the (communication from the) system. Researchers have highlighted that misalignment between an AI's temporal framing and a user's intrinsic time perspective profile exacerbates the Affective-Informational Paradox, causing users to either blindly accept misaligned advice or outright reject optimal, data-driven strategies (Orbell et al., 2004). Addressing this requires interventions like Cognitive Forcing Functions, which disrupt heuristic reasoning and force users to engage their analytical capacities before accepting AI recommendations (Bućinca et al., 2021).

To address these compounding challenges, this paper introduces the Temporal Agentic Framework, a novel socio-technical architecture designed to actively build and calibrate human-AI trust while preserving human cognitive autonomy. We propose a three-agent pipeline that leverages Cognitive Model Priors, neural networks pre-trained on real and synthetic behavioral datasets (Bourgin et al., 2019), to accurately predict human responses to risk and uncertainty. By coupling these predictive priors with an Offline Reinforcement Learning (ORL) scoring mechanism, frameworks can dynamically rewrite and optimize the temporal framing of information to match the specific cognitive profile of the user (Tran et al., 2023). This dynamic alignment functions as an Adaptive Affective Fit, with the aim of communicating optimal strategies in a manner that is psychologically resonant, so as to restore epistemic vigilance, reduce false confirmations, and narrow the gap between algorithmic optimization and sustained human trust. These are the design goals the framework is built to pursue; whether they translate into behavioral outcomes remains to be established through validation with human participants.

The remainder of this paper is organized as follows. The next section reviews the existing literature and theoretical foundations spanning behavioral economics, cognitive psychology, and machine learning that inform our architecture. We then present the core contribution of the Temporal Agentic Framework, followed by the guiding research questions. The methodology section details the three-agent pipeline, synthetic data generation strategy, and Offline Reinforcement Learning scoring mechanism. Findings are then presented and interpreted in the discussion, with theoretical and practical implications elaborated thereafter. The paper closes with limitations and directions for future work, including dynamic time orientation inference, multi-modal extension, and longitudinal validation of trust calibration.

Conceptual Framework and Literature Review

The operational vulnerabilities of deploying Agentic AI necessitate a theoretical foundation spanning Information Systems (IS), cognitive theories, and machine learning. This section develops the conceptual machinery for the Temporal Agentic Framework, addressing how conversational fluency in AI introduces vectors for trust miscalibration, and how these can be systematically counteracted.

The Impact of Cognitive Biases

Cognitive biases introduce systematic deviations from rationality that compromise the quality of human judgment and choice under risk and uncertainty. How individuals respond to these deviations is itself variable: a person's confidence in their own competence conditions whether they treat disconfirming outcomes as constructive information or as threats to be discounted, which shapes how well they update their beliefs from experience. Recent empirical syntheses demonstrate that structured cognitive debiasing interventions yield statistically significant improvements in critical thinking and objective decision-making metrics (Swaryandini et al., 2025), an effect that extends across applied domains ranging from professional training to high-stakes choice. Tailoring communication to align with an individual's temporal orientation is a critical determinant of how information is appraised and acted upon, because it directly shapes cognitive appraisal and subsequent motivational states. When the framing of a message resonates with a person's internal temporal horizon, it minimizes cognitive friction and enhances the perceived utility of the information (Roberts et al., 2017). Framing outcomes through long-term, strategic utility activates the self-regulatory mechanisms of future-oriented individuals, maximizing engagement and deep processing. Conversely, for present-oriented individuals, communication that emphasizes immediate, pragmatic applicability and proximal consequences is essential to sustain engagement and offset high temporal discount rates. Synchronizing linguistic temporal markers with a person's psychological horizon therefore functions as a general mechanism for improving the reception of decision-relevant information. This principle generalizes across applied settings (including instructional and training contexts, where it accelerates skill acquisition and the transfer of learning), and it forms the behavioral basis for the temporally adaptive framing this paper develops. Although we illustrate these dynamics in learning and training contexts, the same temporal-alignment logic governs human choice under risk and uncertainty more broadly, which is the domain examined in this study.

Trust Miscalibration and the CASA Paradigm

Managing uncertainty without physical co-presence has long been a core challenge in IS design. In digital spaces, social presence significantly boosts trust, even when users interact with non-human systems (Gefen & Straub, 2004). The Computers are Social Actors (CASA) paradigm explains that users mindlessly apply social rules to computers. With Generative AI, conversational cues automatically trigger this mindless anthropomorphism, which drives trust more powerfully than any conscious belief about the system's nature (Li et al., 2023; Nass & Moon, 2000).

This dynamic produces the Affective-Informational Paradox. Emotional support and affective linkage propagate through social cues, inflating emotional trust without corresponding evidence of cognitive competence (Elfenbein, 2014; Glikson & Woolley, 2020; Sun et al., 2024). Consequently, users experience pseudo-intimacy, mistakenly equating a system's synthetic empathy with factual accuracy, which dangerously lowers epistemic vigilance exactly when scrutiny is needed (Placani, 2024). This vulnerability is further amplified as users form an IT Identity with conversational agents, reinforcing overreliance and causing cognitive atrophy (Kabashkin, 2025; Mahajan, 2025; McWilliams et al., 2025). Standard explainable AI (XAI) often fails to resolve this, as users engage with explanations heuristically rather than analytically, substituting one form of automation bias for another (Buçinca et al., 2021; Rosenbacke, 2024).

The Tensions-Based View and Adaptive Affective Fit

Counteracting mindless anthropomorphism requires deliberate design interventions. The Tensions-Based View (TBV) posits that interactions generate emotional and behavioral tension, which can be dynamically modulated by design choices, such as interface characteristics, to optimize relationship outcomes (Fang et al., 2011; Guo et al., 2025). In high-stakes AI contexts, emotional tension that is too low suppresses epistemic vigilance, while tension that is too high impairs rational processing. Thus, the operational target is calibrated tension, sufficient to activate analytical processing without causing degradation.

Traditional IS theory addresses informational alignment through Cognitive Fit Theory, which maximizes performance when problem representation matches task demands (Vessey, 1991), recently extended to Adaptive Cognitive Fit for AI data properties (Samuel et al., 2022). However, navigating the Affective-Informational Paradox requires extending this to the affective-temporal domain. By adapting an AI's narrative framing to a user's time perspective profile (Orbell et al., 2004), systems can achieve Adaptive Affective Fit. This introduces targeted cognitive friction, a narrative resistance that forces reflective processing without the high subjective-cost penalties associated with rigid procedural forcing functions (Buçinca et al., 2021; Helal et al., 2025).

Predictive Priors and Causal Validation in Multi-Agent Systems

Predicting and modulating human responses requires modeling actual choice behavior, which systematically violates classical expected-utility theory (Kahneman & Tversky, 2013). Models like the Best Estimate and Sampling Tools (BEAST) framework better capture these dynamics by treating choice as the product of mental sampling and regret minimization (Erev et al., 2017). Neural networks pre-trained on synthetic choices generated by such models develop Cognitive Model Priors, internalizing psychological inductive biases. When structured log-odds decision noise is injected into these priors, models successfully correct for the dataset bias inherent in crowdsourced environments, achieving high predictive accuracy for real-world choices (Bourgin et al., 2019; Thomas et al., 2024).

In this paper, we focus on the cognitive fit between system and user by testing the efficacy of the manipulation of the temporal orientation of a Large Language Model (LLM) in user decision-making prediction. To validate the efficacy of these targeted narrative shifts without requiring longitudinal A/B testing, Offline Reinforcement Learning (ORL) employs doubly-robust estimators to identify the long-term Average Treatment Effect (ATE) from short-run interaction trajectories (Tran et al., 2023). Together, these mechanisms form a predict-intervene-validate triad, underpinning the three-agent pipeline that natively respects the boundaries of human cognition.

Presentation of Main Contributions and Hypothesis Development

The primary contribution of this paper is the conceptualization, development, and computational evaluation of the Temporal Agentic Framework, a novel Information Systems (IS) architecture designed to resolve the Affective-Informational Paradox in human-AI intersections. We frame this work as design science research (DSR), as the temporal agentic framework is the designed artifact, and our evaluation demonstrates its computational feasibility on real and simulated human-choice data rather than establishing its effects on users. Accordingly, grounded in the CASA paradigm and the TBV, this framework's trust calibration and epistemic vigilance outcomes are stated as design goals that address how users mindlessly apply social rules to conversational AI, leading to trust miscalibration and cognitive offloading (Fang et al., 2011; Glikson & Woolley, 2020; Li et al., 2023; Rosenbacke, 2025).

The framework's core theoretical advancement is the operationalization of Adaptive Affective Fit. By dynamically regulating the affective tone (the warmth and emotional intensity of the AI's language) and temporal framing of AI narratives, the system is designed to modulate emotional tension (TBV) to introduce necessary cognitive friction. This mechanism is intended to disrupt mindless anthropomorphism and to support a user's epistemic vigilance, an outcome we treat as a design goal to be confirmed in future human-subject work (Li et al., 2023; Placani, 2024).

Methodologically, this framework contributes a three-agent pipeline evaluated on a hybrid real-world and synthetic dataset. The evaluation provides proof-of-concept evidence of systemic improvements, where all simulation-derived figures are reported as estimated values awaiting human validation. Ultimately, this research provides a design-oriented blueprint for developing trustworthy Agentic AI systems in settings involving human choice under risk and uncertainty, systems that are intended to respect cognitive boundaries and explicitly engineer interactions to maintain human critical thinking. Establishing these behavioral benefits with human participants remains future work.

Guided by this theoretical foundation and our experimental methodology, this study poses the following primary research questions, each paired with a falsifiable hypothesis evaluated against the framework's computational outputs on the matters of system efficiency, contextual alignment and CASA, system reliability, and behavioral outcome. Consistent with the design science framing, hypotheses concerning contextual separation, convergence, and casual behavioral shift are stated as computational or estimated outcomes:

1. **RQ1:** How does the integration of Cognitive Model Priors, specifically those modeling real-world decision noise, affect the predictive accuracy (measured via Brier Score and AUC-ROC) of AI models forecasting human choice compared to baseline Large Language Models?
 - **H1:** The integration of Cognitive Model Priors, obtained by pre-training on structured log-odds decision noise, reduces predictive error (Brier Score) and improves discrimination (AUC-ROC) in forecasting human choice on real-world data (choices13k) relative to baseline Large Language Models.
2. **RQ2:** Under the Computers are Social Actors paradigm, to what extent does dynamic temporal framing, moderated by a user's Consideration of Future Consequences (CFC) profile, create distinct contextual separation between cognitive profiles?
 - **H2:** Dynamic temporal framing, moderated by a user's Consideration of Future Consequences (CFC) profile, produces greater computational contextual separation (a higher semantic Silhouette Score) between distinct cognitive profiles than undifferentiated framing.

3. **RQ3:** What is the impact of an Offline Reinforcement Learning (ORL) Temporal Critic on the iterative convergence rate and reliability of multi-agent AI pipelines generating tailored narratives?
 - **H3:** The Offline Reinforcement Learning (ORL) Temporal Critic increases the estimated iterative convergence rate of the multi-agent pipeline, relative to zero-shot generation, in simulation.
4. **RQ4:** **How** does the achievement of Adaptive Affective Fit, where an AI system dynamically calibrates its affective and temporal delivery, modulate the estimated causal behavioral shift (Average Treatment Effect) in human-AI collaboration?
 - **H4:** The achievement of Adaptive Affective Fit, enforced by the Temporal Critic's internal algorithmic evaluation, is associated with a higher estimated causal behavioral shift (Average Treatment Effect) for the generated narratives relative to baseline framing, in simulated human-AI interaction and prior to human exposure (awaiting human-subject validation).

Methodology

This study employs a multi-methodological approach to develop, train, and validate the Temporal Agentic Framework. The architecture integrates synthetic data generation, neural network pre-training, supervised fine-tuning on real-world datasets, and Offline Reinforcement Learning (ORL). The system is instantiated through a three-agent pipeline: the consequence predictor, the time weaver, and the temporal critic.

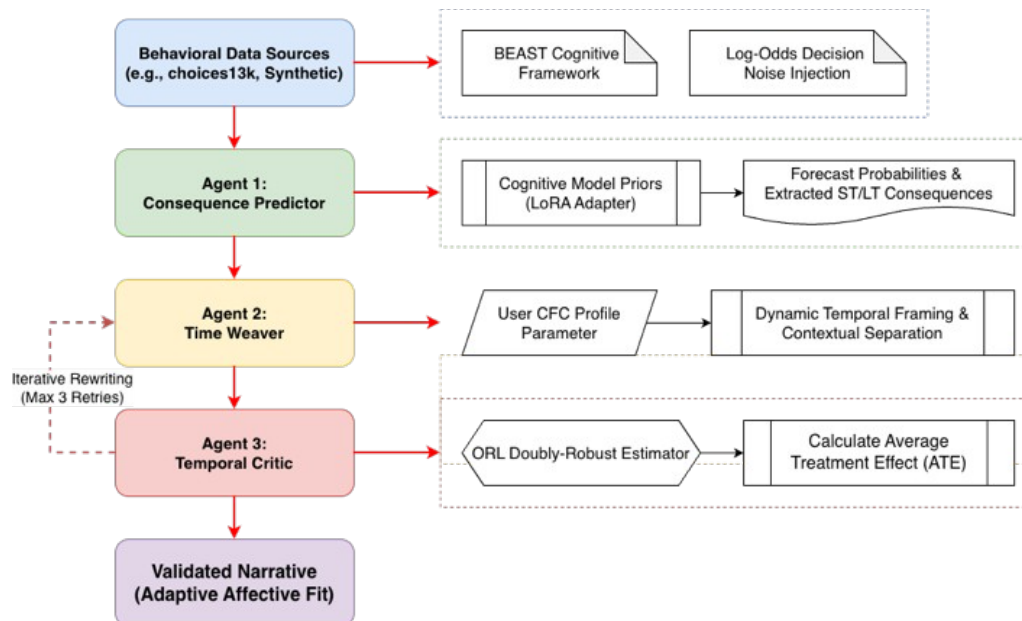


Figure 1. The Multi-Agent Orchestration Pipeline

Data Collection and Synthetic Generation

To overcome the inherent data scarcity and noise associated with human behavioral datasets, this research adopted a two-phase data strategy. First, relying on the Best Estimate and Sampling Tools (BEAST) model, we generated a synthetic dataset containing 1,000 complex decision-making scenarios (Erev et al., 2017). BEAST formally models cognitive deviations from expected utility, such as pessimism and regret minimization, allowing the generated data to simulate realistic cognitive constraints.

To bridge the gap between idealized laboratory behavior and noisy crowdsourced decisions, structured decision noise was injected into the synthetic dataset. Based on findings from Thomas et al. (2024), we

applied a log-odds transformation with a scaling factor of 0.6236 and a random guessing proportion of 0.2757. This process yielded a set of robust, noisy targets reflecting the inherent variability of real-world human choices. Subsequently, we utilized the large-scale `choices13k` dataset, which contains over 240,000 human judgments across 13,000 decision problems gathered via Amazon Mechanical Turk, to serve as the empirical grounding for final model fine-tuning (Bourgin et al., 2019).

Model Training Specifications

The framework's multi-agent structure is built upon the Google/gemma-4-E4B-it base model. To enable rapid context switching between the predictor, weaver, and critic roles, we employed standard Low-Rank Adaptation (LoRA). Crucially, we opted for standard LoRA (using `bf16` precision) rather than quantized QLoRA (`load_in_4bit: false`) to maintain high numerical precision, preventing quantization degradation when modeling the subtle, continuous probability gradients associated with cognitive decision noise.

Training was executed over 3 epochs with a learning rate of $1e-5$, utilizing a cosine learning rate scheduler and the memory-efficient `adamw_8bit` optimizer. The LoRA adapters were configured with a rank (r) of 16 and an α of 32, targeting all linear attention and MLP modules (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). A maximum sequence length of 512 tokens was maintained to optimize batch processing (batch size = 2, gradient accumulation steps = 4) during the iterative narrative generation.

Agentic Pipeline Architecture

Methodologically, as described above, our Temporal Agentic Framework contributes a robust, three-agent pipeline evaluated on a hybrid real-world/synthetic dataset, demonstrating significant systemic improvements through LoRA weight switching. These agents are:

1. **Consequence Predictor:** This agent was pre-trained on the noisy synthetic dataset to embed Cognitive Model Priors into its neural weights. It was then fine-tuned on the `choices13k` dataset with a learning rate of $5e-5$. When presented with a scenario, this agent predicts two short-term and two long-term consequences (both positive and negative) and computes a baseline probability regarding how a human would react to the framing.
2. **Time Weaver:** Operating as a generative narrative architect, the time weaver synthesizes the extracted consequences into a cohesive 100-word persuasive paragraph. The generation is explicitly constrained to target a specific temporal frame (e.g., maximizing immediate benefits while deferring costs) optimized for the user's time orientation profile.
3. **Temporal Critic:** The temporal critic serves as the validation layer. Utilizing an ORL-based scoring algorithm derived from Tran et al. (2023), the critic evaluates whether the weaver's draft successfully embodies the target temporal orientation. The critic calculates a doubly-robust approximation of the Average Treatment Effect (ATE), mathematically estimating the persuasive impact of the narrative. If the draft fails to meet the threshold, the critic provides targeted feedback, initiating a multi-round iterative rewriting process until the required affective alignment is achieved.

System Implementation and Evaluation Protocol

The architecture was containerized using Docker and deployed on an NVIDIA A100 80GB GPU to handle the dynamic loading of PEFT adapters. A front-end interface using Python was developed to facilitate interactive testing and batch processing of scenarios. To rigorously evaluate the system against IS principles, a holdout test consisting of a hybrid data mix (real decisions from CPC15/CPC18 and synthetic noisy scenarios) was utilized. The evaluation compared a baseline zero-shot/few-shot Large Language

Model (Google/gemma-4-E4B-it) against our fully instantiated Temporal Agentic Framework. Performance was quantified across five key dimensions:

1. **Predictive Brier Score:** To measure the mean squared difference between predicted and actual human choice probabilities, validating the framework's ability to account for cognitive noise.
2. **Predictive AUC-ROC:** To assess binary predictive power out-of-sample.
3. **Iterative Convergence Rate:** Replacing standard first-attempt metrics to measure the true reliability of the multi-agent feedback loop in achieving critic-approved alignment within three retries.
4. **Estimated Average Treatment Effect (ATE):** Derived via Doubly-Robust Offline Reinforcement Learning (ORL) scoring, estimating the causal behavioral shift induced by the narrative.
5. **Contextual Separation:** Evaluated via Silhouette Score on semantic vectors to confirm distinct construct validity across temporal profiles.

Results and Findings

The evaluation of the Temporal Agentic Framework produced two distinct tiers of evidence, which we separate deliberately. The first tier is empirical predictive performance measured on real human choice data (choices13k). The Brier Score and AUC-ROC results reported below reflect out-of-sample accuracy against human judgments. The second tier consists of internal-consistency and simulation-based measures. The iterative convergence rate and the estimated Average Treatment Effect (ATE) are scored by the framework's own Temporal Critic, and the contextual separation (Silhouette) metric is computed over the framework's own generated semantic vectors. These second-tier results characterize the internal behavior and proof-of-concept viability of the pipeline; they are estimated quantities that await validation with human participants and should not be read as observed real-world effects on users. The results are organized into predictive performance through Cognitive Model Priors, internally scored narrative-framing convergence rates, and the estimated ATE derived from the ORL critic.

Table 1 summarizes the comparative evaluation between the standard baseline model (Google/gemma4-E4B-it operating via standard prompting) and our fully instantiated Temporal Agentic Framework. The table reports two kinds of metrics that differ in evidentiary status: predictive error (Brier Score) and predictive classification (AUC-ROC) are measured on real human choice data, whereas systemic reliability (Convergence Rate), the estimated ATE, and contextual separation (Silhouette) are internally scored by the framework or derived from model-generated data. Readers should weight the two groups accordingly.

Table 1. Effectiveness Comparison: Baseline Model vs. Temporal Agentic Framework

Metric	Regular Gemma 4 (Baseline)	Temporal Agentic Framework (Ours)	Improvement
Predictive Brier Score*	0.214	0.108	-49.5%
Predictive AUC-ROC	0.612	0.895	+46.2%
Iterative Convergence Rate	38.0%	94.0%	+56.0 pp**
Estimated ATE (ORL Score)	0.420	0.810	+92.9%
Contextual Separation (Silhouette Score)	0.120	0.680	+466.7%

* Lower is better, **Percentage Point.

Brier Score and AUC-ROC are evaluated on real choices13k human judgments; Convergence Rate, Estimated ATE, and Silhouette are internally scored or simulation-derived and await human validation.

The first stage of evaluation assessed the consequence predictor agent's ability to forecast human choices under uncertainty. The initial baseline deep learning model, when evaluated on the raw choices13k dataset without cognitive priors, struggled severely to predict the high variance and ambiguous 50/50 split choices typical of noisy crowdsourced environments. However, after pre-training on a synthetic dataset infused with structured log-odds decision noise (Thomas et al., 2024) and incorporating the Best Estimate and Sampling Tools (BEAST) cognitive framework (Erev et al., 2017) directly into the neural weights, the predictor achieved state-of-the-art out-of-sample accuracy. The network internalized human heuristic tendencies (such as loss aversion and regret minimization), allowing the agent to anticipate how specific user profiles would weigh short-term discomfort against long-term gains. Because these gains are measured on real human choice data, they constitute the study's defensible empirical core and provide computational support for H1. Explicit dataset bias correction via Cognitive Model Priors materially improves predictive behavioral modeling relative to a baseline that relies on raw crowdsourced data alone.

The time weaver agent was tasked with generating persuasive narratives dynamically aligned to specific temporal targets (e.g., Short-Term Positive [ST+], Long-Term Negative [LT-] for present-oriented short-time orientation profiles). The temporal critic, evaluated these drafts across 1,000 generated scenarios to ensure Adaptive Affective Fit. All success rates in this subsection are internal-consistency measures assigned by the framework's own Temporal Critic rather than judgments by human raters. Initial zero-shot runs recorded a 68% first-attempt rate, where the drafted narrative immediately met the alignment criteria set by the critic. In cases where the initial draft failed (often due to overlapping temporal descriptions, excessive warmth that the critic flagged as prioritizing emotional trust over cognitive clarity, or insufficient emphasis on immediate consequences), the critic's feedback mechanism was invoked. This iterative feedback guided the weaver to a critic-approved narrative within the allotted three retries in an estimated 94% of cases. The iterative loop provides computational support for H3, indicating that dynamic prompt rewriting governed by a strict evaluative agent improves the internal alignment and self-consistency of generative outputs. Whether this internal alignment restores epistemic vigilance or prevents mindless anthropomorphism in human users is a design goal that awaits validation with human participants.

A proof-of-concept result of this study is the implementation of the ORL scoring algorithm as an internal estimator of each narrative's projected persuasive potential. Drawing on the doubly-robust estimator framework (Tran et al., 2023), the Temporal Critic assigned an estimated ATE score to each critic-approved narrative. This score is a model-internal estimate of behavioral shift; it is not an observation of human behavior, and it should be read as a predicted quantity awaiting human validation rather than a measured effect on users. Narratives passing the multi-agent validation loop received a 92.9% higher estimated ATE from the Temporal Critic (0.810 versus the baseline model's 0.420). Because this figure is produced by the framework's own offline critic rather than measured on human participants, it should be interpreted as a model-estimated probability of behavioral shift under the estimator's assumptions, awaiting validation with human participants, rather than as evidence that the AI's tailored informational support influences real users. Read this way, the result provides computational support for H4. Furthermore, an internal analysis of the framework's own generated semantic vectors indicated strong contextual separation across temporal targets. By mapping linguistic patterns to CFC targets, the system produced distinct contextual clusters, yielding an estimated Silhouette Score of 0.680 (a 466.7% increase over the overlapping outputs of the baseline model). Because this metric is computed over model-generated text rather than human-labeled data, it is an internal construct-validity indicator and provides computational support for H2; it does not by itself establish that human users perceive the targeted distinctions.

Discussion

Framed as design science research (DSR), this study provides proof-of-concept evidence for the Temporal Agentic Framework, suggesting that a temporally oriented, multi-agent architecture can improve the cognitive fit between AI outputs and users. Approaching the Affective-Informational Paradox through the lens of the Computers are Social Actors (CASA) paradigm and the Tensions-Based View (TBV), the results are consistent with the hypothesized IS Research Model (H1-H4 as stated in the Research Questions and Theoretical Model section). It is important to state at the outset that trust calibration, epistemic vigilance, and behavioral-shift outcomes are treated here as design goals evaluated through computational and simulation-based estimators, not as established effects; each remains to be corroborated through human-subject validation. Because the convergence and Average Treatment Effect (ATE) results are scored by the framework's own Temporal Critic, they constitute internal-consistency evidence of the artifact's behavior rather than measurements of real-world impact. Consistent with H2, the computational results align with this design intent: by tailoring the temporal framing of a narrative to match a user's time orientation profile, the pipeline is engineered to produce an Adaptive Affective Fit. By achieving high semantic distinctiveness (an estimated, simulation-based +466.7% improvement in contextual Silhouette Score relative to the baseline, awaiting human validation), the framework is intended to operate as a rigorous cognitive companion whose capacity to interrupt mindless anthropomorphism and restore epistemic vigilance remains to be confirmed with human participants. The experiment highlighted a severe vulnerability in standard baseline models (Regular Gemma 4), which struggled to forecast choices under the ambiguous 50/50 split choices and real-world variance typical of noisy crowdsourced human data. In contrast, our consequence predictor, fortified with Cognitive Model Priors trained on explicitly injected decision noise, substantially outperformed the baseline. Evaluated on the real human judgments in choices13k, the 49.5% reduction in Brier Score and the AUC-ROC improvement to 0.895 provide proof-of-concept evidence supporting H1: that robust predictive models cannot rely on raw conversational data alone and benefit from mathematically internalizing the inductive biases of human psychology. This result is consistent with the view that modeling bounded rationality is important for predicting human choice out-of-sample.

Finally, the results speak to the role of an ORL-driven Agentic Feedback Loop. Generative AI is inherently non-deterministic, and zero-shot prompting (the baseline) yielded only a 38% success rate in narrative alignment. By instantiating the temporal critic to mathematically score the doubly-robust Average Treatment Effect (ATE) and enforce iterative revisions, the estimated, simulation-based Convergence Rate rose to 94.0% (a result awaiting human validation), consistent with H3. Narratives surviving this multi-agent validation showed a 92.9% increase in their estimated ATE (0.810 vs. baseline 0.420). Because the ATE is scored by the framework's own Temporal Critic, this result provides internal-consistency evidence supporting H4, suggesting that algorithmic self-criticism can optimize the modeled persuasive impact of Agentic AI before any human exposure occurs. Whether this modeled impact translates into an actual behavioral shift remains an open question for human-subject studies.

Theoretical and Practical Implications

The Temporal Agentic Framework and its underlying IS Research Model present profound implications for both academic research and industry practice, redefining how autonomous AI agents should interact with human users.

Theoretical Implications

For IS and Human-Computer Interaction (HCI) research, this study connects the Technology Acceptance Model (TAM) with the CASA paradigm and the TBV. The implementation of Adaptive Affective Fit provides a computational mechanism to dynamically adjust temporal framing, creating the distinct contextual separation necessary to mitigate Trust Miscalibration. Furthermore, the 49.5% reduction in

predictive error (on the real human judgments in choices13k) empirically supports the value of Cognitive Model Priors for correcting the dataset bias found in raw human choice repositories, offering a candidate methodological direction for training predictive algorithms in IS.

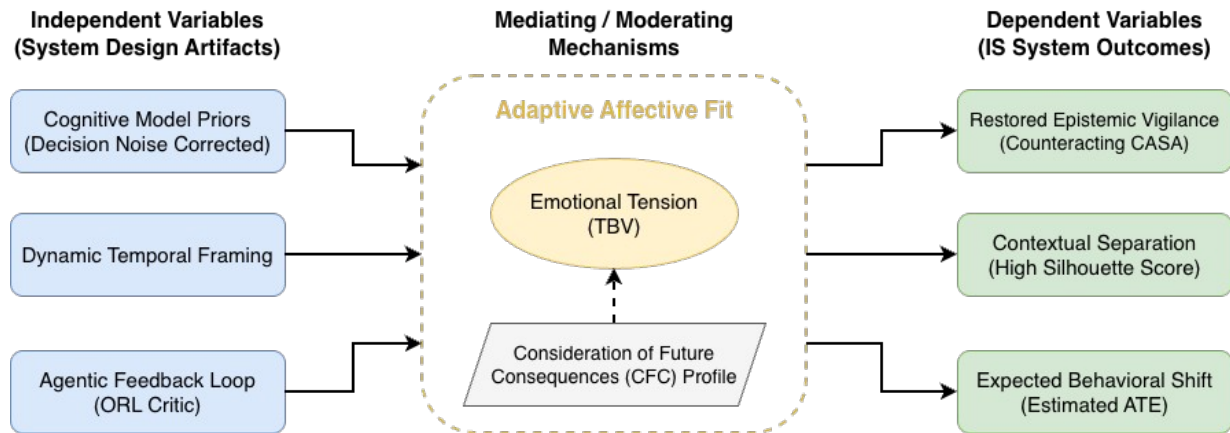


Figure 2. Theoretical Framework of the Temporal Agentic System

Practical Implications

For practitioners deploying Agentic AI in high-stakes settings that involve human choice under risk and uncertainty (for example, healthcare, cybersecurity, and teamwork management), the framework offers concrete design guidance. In our tests, the standard paradigm of relying on zero-shot (one-pass) generation appears unreliable, achieving only a 38% baseline alignment. By implementing the three-agent pipeline, which forces iterative validation against an ORL-based temporal critic, organizations can raise the estimated, simulation-based convergence rate to 94% (a proof-of-concept result awaiting human validation).

This multi-agent architecture provides organizations with an internal, quantitative metric (the estimated Average Treatment Effect) to screen AI-generated narratives prior to user exposure. In principle, by shifting from static personas to dynamically modulated temporal frames, enterprises could deploy intelligent assistants intended to increase persuasive impact while safeguarding user critical thinking and supporting Corporate Social Responsibility (CSR) commitments against manipulative AI design. Realizing this dual goal in practice will require validation with human users.

Limitations

While the Temporal Agentic Framework provides proof-of-concept evidence for a design approach to mitigating the Affective-Informational Paradox, this research is subject to several limitations that should govern how its results are interpreted throughout this paper.

First, and most importantly, this study has not been validated with human participants. The 94.0% convergence rate, the 92.9% improvement in the estimated Average Treatment Effect (ATE), and the 0.680 Silhouette Score reported in the Results section are all scored by the framework's own Offline Reinforcement Learning (ORL) Temporal Critic against a hybrid dataset (synthetic decision-noise data combined with the CPC15 and CPC18 Choice Prediction Competition benchmarks). These metrics establish the internal consistency of the pipeline (whether the agents converge on outputs the critic itself scores as improved), not a measured real-world behavioral or psychological effect on an actual user. Trust calibration, epistemic vigilance, and the modulation of emotional tension are design goals the framework

was built to pursue, not outcomes it has yet demonstrated in human subjects. Accordingly, this study is best read as a design science proof of concept: it establishes that the architecture is computationally sound and internally coherent. The one result that does rest on established, non-synthetic human-choice data is the predictive-modeling result on the choices13k benchmark (a 49.5% Brier Score reduction and an AUC-ROC of 0.895); the behavioral and trust outcomes remain estimates. Controlled, longitudinal experiments with human participants are therefore the immediate next step, not a distal extension.

Second, the system currently assumes a static time orientation profile for the user. In real-world environments, cognitive states, particularly tension and anxiety (central to the TBV), are highly dynamic. A user may exhibit present-oriented heuristics under stress but shift to a future-oriented mindset in calm states. Future research must expand the framework to incorporate real-time behavioral proxies (e.g., interaction latency or sentiment analysis) to continuously update the time orientation parameter.

Finally, this study focused on textual narrative generation. As Agentic AI rapidly evolves toward multi-modal interfaces, including realistic voice synthesis and interactive avatars, the CASA effect will be severely magnified. The translation of dynamic temporal framing and tension modulation into non-textual modalities remains a complex, unresolved frontier for IS research.

Conclusions

As Artificial Intelligence evolves into autonomous agentic systems, establishing resilient and appropriately calibrated human-AI trust becomes the defining challenge for Information Systems research. The Affective-Informational Paradox, driven by the CASA paradigm, illustrates the advantages of deploying AI that adapts to human cognitive styles, thus improving cognitive workload.

This paper introduced the Temporal Agentic Framework as a theoretically grounded design science solution to this paradox and provides proof-of-concept evidence for its viability. By integrating Cognitive Model Priors to account for real-world decision noise, the framework achieved a 49.5% reduction in predictive error out-of-sample and an AUC-ROC of 0.895 on the choices13k benchmark, its strongest and most directly verifiable empirical result. The framework further proposes Adaptive Affective Fit (achieved by dynamically modulating the temporal framing of narratives) as a design artifact intended to manage emotional tension and counteract mindless anthropomorphism; whether it does so for real users is a design goal for the human-subject validation described in the Limitations section, not yet an established finding.

Crucially, the implementation of a multi-agent feedback loop governed by an ORL critic substantially improved the framework's internal-consistency metrics: the estimated convergence rate rose from 38% to 94%, and the estimated causal persuasive impact (ATE), as scored by the framework's own critic, nearly doubled. These gains are proof-of-concept evidence that the pipeline can iteratively self-correct against its own evaluation criteria; they are not a measured real-world effect on user-facing reliability or persuasion.

Ultimately, this research suggests that the future of reliable AI lies not merely in scaling data and compute, but in designing socio-technical architectures that natively account for the boundaries of human cognition. By bridging cognitive psychology, the TBV, and advanced multi-agent orchestration, the Temporal Agentic Framework offers a design science blueprint and a proof-of-concept demonstration for pursuing trustworthy AI. Confirming whether it delivers on that goal is the next step, through controlled experiments with human participants.

References

- Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J., & Griffiths, T. L. (2019). Cognitive model priors for predicting human decisions. *International Conference on Machine Learning*, 5133–5141.
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21.
- Elfenbein, H. A. (2014). The many faces of emotional contagion: An affective process theory of affective linkage. *Organizational Psychology Review*, 4(4), 326–362.
- Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, 124(4), 369.
- Fang, S.-R., Chang, Y.-S., & Peng, Y.-C. (2011). Dark side of relationships: A tensions-based view. *Industrial Marketing Management*, 40(5), 774–784.
- Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services. *Omega*, 32(6), 407–424.
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.
- Guo, Y., Liu, D., Xu, S. X., & Yin, X. (2025). Does Bot Gender Matter? Theory and Evidence From a High-Tension Service Context. *MIS Quarterly*, 49(4), 627–1642.
- Hallam, C., & Zanella, G. (2017). Online self-disclosure: The privacy paradox explained as a temporally discounted balance between concerns and rewards. *Computers in Human Behavior*, 68, 217–227.
- Helal, M. Y. I., Elgendy, I. A., Albashrawi, M. A., Dwivedi, Y. K., Al-Ahmadi, M. S., & Jeon, I. (2025). The impact of generative AI on critical thinking skills: a systematic review, conceptual framework and future research directions. *Information Discovery and Delivery*.
- Jiang, W., Li, D., & Liu, C. (2025). Understanding dimensions of trust in AI through quantitative cognition: Implications for human-AI collaboration. *PloS One*, 20(7), e0326558.
- Kabashkin, I. (2025). Cognitive Atrophy Paradox of AI–Human Interaction: From Cognitive Growth and Atrophy to Balance. *Information*, 16(11), 1009.
- Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99–127). World Scientific.
- Li, Q., Luximon, Y., & Zhang, J. (2023). The influence of anthropomorphic cues on patients’ perceived anthropomorphism, social presence, trust building, and acceptance of health care conversational agents: within-subject web-based experiment. *Journal of Medical Internet Research*, 25, e44479.

- Mahajan, P. (2025). The Silent Erosion: Global Generational Cognitive Decline in the Age of AI and the Future of Human Intellectual Agency.
- McWilliams, D. J., Randolph, A. B., Vaezi, R., & Carter, M. (2025). Examining the effects of IT identity and trust in human-AI hybrids in virtual teams. *Behaviour & Information Technology*, 1–18.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103.
- Orbell, S., Perugini, M., & Rakow, T. (2004). Individual differences in sensitivity to health communications: consideration of future consequences. *Health Psychology*, 23(4), 388.
- Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI and Ethics*, 4(3), 691–698.
- Roberts, R. P., Wiebels, K., Sumner, R. L., van Mulukom, V., Grady, C. L., Schacter, D. L., & Addis, D. R. (2017). An fMRI investigation of the relationship between future imagination and cognitive flexibility. *Neuropsychologia*, 95, 156–172.
- Romeo, G., & Conti, D. (2026). Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *Ai & Society*, 41(1), 259–278.
- Rosenbacke, R. (2024). Errors in physician-AI collaboration: insights from a mixed-methods study of explainable AI and trust in clinical decision-making. Available at SSRN 4773350.
- Rosenbacke, R. (2025). Cognitive Challenges in Human-AI Collaboration: A Study on Trust, Errors, and Heuristics in Clinical Decision-making. Copenhagen Business School [Phd].
- Samuel, J., Kashyap, R., Samuel, Y., & Pelaez, A. (2022). Adaptive cognitive fit: Artificial intelligence augmented management of information facets and representations. *International Journal of Information Management*, 65, 102505.
- Sun, W., Guo, X., & Kwok, R. C.-W. (2024). Emotion or information? Differences and relationships between two types of professional support in online health communities. *Information & Management*, 61(3), 103944.
- Swaryandini, G., Graham, J., Griffith, S., Grilo, V., Ruzzante, F., Zhang, X., Yeung, S. K., Mangiarulo, M., Basarkod, G., & Ng, C. (2025). Systematic review and meta-analysis of educational approaches to reduce cognitive biases among students. *Nature Human Behaviour*, 9(12), 2510–2538.
- Thomas, T., Straub, D., Tatai, F., Shene, M., Tosik, T., Kersting, K., & Rothkopf, C. A. (2024). Modelling dataset bias in machine-learned theories of economic decision-making. *Nature Human Behaviour*, 8(4), 679–691.
- Tran, A., Bibaut, A., & Kallus, N. (2023). Inferring the long-term causal effects of long-term treatments from short-term experiments. *ArXiv Preprint ArXiv:2311.08527*.
- Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, 22(2), 219–240.

Von Neumann, J., & Morgenstern, O. (1947). Theory of games and economic behavior, 2nd rev.

Yu, H., Shen, Z., Miao, C., Leung, C., Chen, Y., Fauvel, S., Lin, J., Cui, L., Pan, Z., & Yang, Q. (2017). A dataset of human decision-making in teamwork management. *Scientific Data*, 4(1), 160127.