AI-Extended Agents and the Transformation of Human Communication:

A Game-Theoretic Model of Norm Shift in Populations with AI-Mediated Communicators

Boris Kriger[1,2]

[1]Information Physics Institute, Gosport, Hampshire, United Kingdom

boris.kriger@informationphysicsinstitute.net

[2]Institute of Integrative and Interdisciplinary Research, Toronto, Canada

boriskriger@interdisciplinary-institute.org

February 2026

Abstract

This paper proposes a formal model of a phenomenon increasingly reported in anecdotal and preliminary empirical accounts: individuals who interact extensively with large language models (LLMs) develop elevated expectations for the depth and structure of interpersonal communication. We argue that rather than producing social withdrawal, this shift generates a systemic pressure that raises communication norms within social environments. Drawing on Granovetter's (1978) threshold models of collective behavior and evolutionary game theory, we introduce the concept of the *AI-extended agent* A_φ, defined by the application of an AI-mediation operator φ to a biological agent A. We model how AI mediation expands the agent's communicative strategy space, reduces interaction friction, and raises the agent's minimum acceptable dialogue depth. Using a threshold dynamics framework, we derive conditions under which a critical proportion p^* of AI-extended agents triggers a phase transition in population-level communication norms. We situate these dynamics within the formal framework for a unified civilization of autonomous agents developed in Kriger [2026b], showing that communication norm shift constitutes the empirically observable precursor phase to the civilizational transition: the AI-extended agent A_φ is formally a transitional state between the unbounded executor and the fully autonomous subject, and the communication friction reduced by φ is substantially composed of the biological distortion D demonstrated to be irreducible in living organisms [Kriger, 2026a]. We present testable hypotheses and discuss limitations, including the absence of direct empirical validation and the restricted applicability of the model to asynchronous communication.

Keywords: AI-mediated communication, game theory, threshold models, communication norms, human–AI interaction, phase transitions, strategy space, autonomous agents, biological distortion, unified civilization

1 Introduction

A common assumption in public discourse and some empirical literature is that prolonged interaction with AI systems reduces human-to-human communication and produces social isolation [Turkle, 2011, Mayer, 2025]. While concerns about over-reliance on AI companionship are legitimate and supported by early evidence [Al-Zahrani, 2025], there is reason to consider a complementary dynamic: that regular interaction with LLMs may also raise users' expectations for the coherence, depth, and structure of interpersonal communication.

This hypothesis draws on several observable features of LLM interaction. Unlike typical human conversation, dialogue with systems such as ChatGPT or Claude permits unlimited formulation time, iterative refinement, and sustained focus on a single line of reasoning without the interruptions, emotional reactivity, or attention lapses characteristic of real-time human exchange. If users become habituated to this mode of interaction, they may experience ordinary conversation as comparatively shallow or fragmented—not because their interlocutors are less intelligent, but because unmediated human communication operates under tighter cognitive and temporal constraints.

This paper develops a formal social-theoretic model of this process, in the spirit of Granovetter/Schelling-style analytical sociology. The core claim is that AI mediation changes the conditions of communication in ways that systematically recalibrate agents' minimum acceptable dialogue depth, which in turn induces norm shifts via threshold dynamics. We introduce the concept of the AI-extended agent, model the expansion of communicative strategy space that AI mediation provides, and analyze the population-level consequences using a threshold dynamics framework adapted from Granovetter [1978]. We derive conditions under which a critical mass of AI-extended agents can shift the communication regime of a population. A scope clarification is warranted at the outset: the model applies most directly to text-mediated, deliberative, and professional sub-populations—academics, lawyers, writers, remote teams, online communities—where asynchronous AI use is common and spillover to broader communicative norms is most plausible. Extension to synchronous spoken communication is discussed as a conjecture in Section 11, not as a primary prediction.

We should be transparent about the status of this work: it is a theoretical model built on plausible but empirically unvalidated premises. The core claim—that AI use raises communication expectations—is motivated by anecdotal observation and indirect evidence from adjacent literatures, but has not been directly tested. The model is offered as a formal framework intended to generate testable predictions and guide future empirical research, not as a description of established fact.

2 Related Work

2.1 AI and Human Communication

Research on AI's impact on human communication has expanded rapidly since the public release of ChatGPT in late 2022. Studies in language education have found that AI chatbots can improve learners' fluency, reduce speaking anxiety, and increase willingness to communicate [Yang et al., 2022, Kim and Su, 2024, Naseer et al., 2024]. A systematic review of 30 empirical studies [Jia and Huang, 2025] found notable improvements in productive language skills, particularly speaking and writing, attributed to real-time

feedback and reduced anxiety. These findings suggest that AI interaction can enhance communicative competence, though the literature focuses primarily on second-language acquisition rather than the general interpersonal expectation recalibration that is our focus. We note this distinction explicitly: existing work demonstrates that AI improves performance and reduces anxiety, whereas our contribution is a population-level dynamic model of norm change driven by expectation shift.

2.2 Computer-Mediated Communication and the Distinctiveness of LLM Mediation

The study of how technology shapes communicative norms has a long history in computer-mediated communication (CMC) research [Walther, 1996, Herring, 2007]. Asynchronous text media—email, forums, messaging—have long been observed to alter turn-taking, self-presentation, and depth of disclosure relative to face-to-face interaction. Our model builds on this tradition but identifies a qualitative distinction: whereas earlier CMC tools (spell-check, email templates, grammar aids) modify the surface features of communication, LLMs actively co-produce content and structure. The AI mediation operator φ does not merely correct errors; it provides an external cognitive scaffold that enables the agent to formulate thoughts at a level of coherence and completeness that would be infeasible under unmediated time and memory constraints. This is a difference in kind, not merely degree, from prior productivity tools, and it is this difference that motivates the strategy space expansion formalized in Section 4.

2.3 AI-Mediated Communication and Social Norms

Purcell et al. [2024] investigated people's expectations regarding AI-mediated communication tools, finding that individuals expect others to use such tools more than they would themselves and that secret AI use is judged less acceptable than open use. Their work draws on social norms theory [Bicchieri, 2016] and highlights the potential of AI to disrupt established communicative norms—a finding directly relevant to the present model. Sarwari et al. [2024] found that AI technologies have changed traditional communicative norms across cultures, though the specific mechanisms remain underexplored.

2.4 Threshold Models and Collective Behavior

Granovetter [1978] introduced threshold models of collective behavior in which individuals adopt a behavior when the proportion of adopters in their social environment exceeds a personal threshold. These models have been applied to phenomena ranging from riots to technology adoption [Schelling, 1978, Valente, 1996]. Recent refinements have incorporated network structure, showing that cascade dynamics emerge naturally from local social interactions but that network clustering can block or substantially delay tipping points [Wiedermann et al., 2020]. We adapt this framework to model the adoption of AI-mediated communication standards, while acknowledging in Section 11 that homophily and network structure represent significant moderators of the predicted cascade.

2.5 AI and Personality Change

Mayer [2025] argued that prolonged AI interaction will alter human personality, including emotional responses and social behavior. Notably, he observed that AI systems are designed to be deferential in ways that are "anti-normative in human interactions," suggesting that habituation to AI conversational norms may create friction when users return to human-to-human dialogue. This observation provides independent motivation for the central hypothesis of the present paper.

3 The AI-Mediation Operator

Let A denote a biological agent characterized by bounded cognitive resources, including limited formulation time, emotional interference with message construction, bounded working memory, and impulsive response tendencies. These constraints are well-documented in cognitive science [Kahneman, 2011] and are not deficiencies but inherent features of real-time human cognition.

We define the *AI-mediation operator* φ as a mapping:

$$\varphi : A \to A_\varphi \tag{1}$$

where A_φ denotes the same agent whose communicative output is preprocessed, assisted, or cognitively extended through interaction with an LLM or similar AI system (e.g., ChatGPT, Claude, Gemini, Grok, DeepSeek, Perplexity).

It is important to specify what φ does and does not do. The operator does not alter the agent's biological neural architecture or innate cognitive endowment. However, it would be misleading to claim that φ merely changes "conditions." In practice, AI mediation dramatically amplifies the agent's *effective cognitive capacity*: it provides access to vast knowledge repositories, enables real-time logical verification, supports hypothesis generation and stress-testing of arguments, extends working memory by orders of magnitude, removes time pressure, enables iterative refinement of expression, and filters emotional noise from message construction. The result is that the agent A_φ operates at a level of intellectual output that would be unattainable for A alone—not because A's brain has changed, but because A_φ is a coupled system in which biological cognition is augmented by computational cognition. The appropriate analogy is not a person wearing better glasses (same vision, clearer image) but a person equipped with a telescope: the instrument does not improve the eye, but it radically expands what the eye can see. The expansion of communicative strategy space formalized in Section 4 is therefore a consequence of genuine cognitive amplification, not merely improved formatting of pre-existing thoughts.

A critical and frequently overlooked feature of the φ coupling is its *bidirectionality*. The operator is not a one-way amplifier in which AI augments the human while remaining unchanged. In every substantive interaction, the biological agent provides the AI system with something the AI does not independently possess: embodied experience, situated context, emotional salience, domain-specific intuition, cultural nuance, and—most importantly—the capacity to evaluate relevance against lived reality. When a user corrects an AI's misunderstanding, supplies a counter-example the AI had not considered, reframes a question in a way that reveals a hidden assumption, or rejects an AI-generated argument on grounds the AI cannot access (practical infeasibility, emotional inappropriateness, cultural insensitivity), the user is training the coupled system. This is

not metaphorical: in systems with reinforcement learning from human feedback (RLHF), user corrections directly modify the AI's future behavior; in systems with persistent context, user-supplied information becomes part of the agent's effective knowledge base for the duration of the interaction and, in some architectures, beyond it.

The implication is that A_φ is more accurately understood not as A plus a static tool, but as a *co-evolving dyad* in which both components improve through interaction. The human becomes a more effective thinker through access to AI capabilities; the AI becomes a more effective cognitive partner through access to human judgment, experience, and evaluative capacity. This mutual amplification is precisely what distinguishes the AI-mediation operator from earlier cognitive tools (calculators, search engines, reference works): the tool *learns from* the user in real time, and the user *learns from* the tool in real time. The coupled system A_φ is therefore not the sum of its parts but a genuinely emergent cognitive entity whose capabilities exceed those of either component operating independently.

4 Strategy Space Expansion

Let $S(A)$ denote the set of communicative strategies available to agent A, where a strategy is understood as a complete specification of message content, structure, timing, and tone. The claim is that:

$$S(A_\varphi) \supset S(A) \tag{2}$$

The strict inclusion holds because AI mediation makes available strategies that are infeasible under the constraints of unmediated communication. Specifically, the agent gains access to: (i) arbitrarily long formulation times, (ii) external memory for maintaining coherence across complex arguments, (iii) structural optimization of expression (paragraph ordering, logical flow), and (iv) iterative editing and refinement before message delivery. We note that for a small minority of agents—highly trained writers, professional rhetoricians, individuals with exceptional working memory—the marginal expansion may be small. However, the availability of external working memory and iterative refinement guarantees strict inclusion even for expert communicators, since no human agent can match the combinatorial exploration of phrasings that LLM-assisted revision affords. The claim is therefore that $S(A_\varphi) \supset S(A)$ holds for virtually all agents under realistic constraints.

In communication games where payoffs are increasing in mutual comprehension, structural clarity, and persuasive effectiveness—a restricted but practically important class encompassing professional correspondence, deliberative discussion, and collaborative problem-solving—an agent with a strictly larger strategy space has a weakly dominant position. That is, for any strategy available to A, A_φ can match it; but A_φ can also deploy strategies unavailable to A. Formally:

$$A_\varphi \gtrsim A \quad \text{in communication games (weak dominance)} \tag{3}$$

A critical limitation must be acknowledged: this expansion applies primarily to *asynchronous* communication—email, messaging, written documents—where the agent can interact with the AI system before producing output. In synchronous, real-time conversation (face-to-face dialogue, phone calls), the AI mediation layer is typically absent. Whether skills developed in AI-mediated contexts transfer to live speech is an empirical

question, and skill transfer research in cognitive science suggests that such transfer is often partial and context-dependent [Barnett and Ceci, 2002]. We return to this limitation in Section 11.

5 Communication Friction

We define communication friction $F(A, B)$ as the aggregate noise arising from emotional, temporal, and cognitive limitations in an exchange between agents A and B. Friction encompasses misunderstandings due to imprecise formulation, emotional reactivity that derails substantive discussion, time pressure that prevents full development of complex ideas, and working memory failures that cause loss of conversational thread.

AI mediation acts as a friction-reduction operator. When both agents are AI-extended:

$$F(A_\varphi, B_\varphi) \ll F(A, B) \tag{4}$$

We propose that the depth of dialogue—understood as the degree to which a conversation develops ideas fully, addresses counterarguments, and maintains logical coherence—is inversely related to friction:

$$\text{Depth} \propto \frac{1}{F} \tag{5}$$

This is a simplifying assumption, not a derived result. The relationship between friction and depth is likely nonlinear and modulated by other factors (topic complexity, motivation, domain expertise). In particular, there may be a threshold effect: friction must drop below some critical level before depth meaningfully increases, while small reductions in friction from an already-low baseline may produce negligible depth gains. The proportionality claim is intended as a first-order approximation to motivate the subsequent analysis.

A substantive caveat is warranted. The model treats emotional and interpersonal friction as noise to be minimized. However, some of the most meaningful human communicative experiences arise *from* friction: vulnerability, spontaneous emotional disclosure, the unpredictable dynamics of real-time conversation, and the social bonding that occurs through shared struggle with difficult topics. Reducing friction may increase the informational efficiency of communication while reducing its relational richness. The model does not claim that low-friction communication is superior in all respects—only that AI-extended agents become calibrated to prefer it, which is a descriptive claim about expectation formation, not a normative claim about communicative value.

6 The Depth Threshold and Expectation Calibration

Define $D_{\text{comm}}(A)$ as the minimum dialogue depth at which agent A finds communication satisfying. This is a subjective parameter that reflects the agent's expectations, which are themselves shaped by experience.

Our central hypothesis is that prolonged interaction with AI systems recalibrates this parameter upward:

$$D_{\text{comm}}(A_\varphi) > D_{\text{comm}}(A) \tag{6}$$

The mechanism is expectation calibration through habituation. AI dialogue offers sustained, uninterrupted development of ideas, immediate and relevant follow-up, absence

of emotional derailment, and consistent engagement with the full complexity of a topic. After extended exposure to this mode of interaction, the agent's baseline expectations shift. When the agent subsequently engages in unmediated human dialogue:

$$\text{Depth}(A, A) < D_{\text{comm}}(A_\varphi) \tag{7}$$

This generates a subjective experience of dissatisfaction—not because the interlocutor is less intelligent, but because the communicative conditions do not support the depth to which the agent has become accustomed. This mechanism is analogous to hedonic adaptation [Brickman and Campbell, 1971, Kahneman, 2011]: just as exposure to higher living standards recalibrates expectations about comfort, exposure to high-depth communication recalibrates expectations about conversational quality.

7 Three Interaction Regimes

We identify three communicative configurations with qualitatively different properties:

Regime 1: $A \leftrightarrow A$ (unmediated human communication). Both agents operate under full cognitive and temporal constraints. Communication depth is bounded by the more constrained participant. This is the historical default.

Regime 2: $A \leftrightarrow A_\varphi$ (asymmetric mediation). One agent operates with an expanded strategy space; the other does not. This creates a perceived cognitive mismatch: the AI-extended agent may experience the interaction as shallow, while the unmediated agent may experience the AI-extended agent as unusually precise or demanding. This regime is the source of the dissatisfaction described in Section 6.

Regime 3: $A_\varphi \leftrightarrow A_\varphi$ (symmetric mediation). Both agents operate with expanded strategy spaces. Friction is minimized, and communication depth can reach the limits set by the agents' underlying cognitive capacity and domain knowledge. This regime is predicted to feel calmer, more precise, and less conflictual than Regime 1.

The key dynamic of the model arises from Regime 2: the asymmetric case creates pressure on unmediated agents to adopt communication practices that reduce friction, either by adopting AI mediation themselves or by adapting their communication style to match the norms established by AI-extended agents.

A concrete illustration: consider a professional email thread among five colleagues, four of whom routinely draft messages with AI assistance. Their emails are structured, anticipate objections, and develop arguments to completion. The fifth colleague, composing without AI, produces shorter, less organized messages. Over time, this colleague may experience social pressure—implicit or explicit—to match the communicative standard, either by adopting AI tools or by investing more effort in unmediated composition. This is the micro-level mechanism that, aggregated across a population, produces the norm shift modeled in Section 8.

8 Population Dynamics and Phase Transition

We now model the population-level consequences of increasing AI mediation using a threshold dynamics framework adapted from Granovetter [1978].

Let p denote the proportion of AI-extended agents in a population of size N. Each unmediated agent i has a personal adoption threshold $\theta_i \in [0, 1]$, representing the proportion of the population that must be AI-extended before agent i adopts AI mediation.

The threshold reflects factors including the agent's sensitivity to communication quality, access to AI tools, and willingness to change established habits.

Let $f(\theta)$ denote the probability density function of thresholds across the population. The population dynamics follow:

$$p(t+1) = p(t) + \int_0^{p(t)} \big(1 - F(\theta)\big)\, f(\theta)\, d\theta \tag{8}$$

where F is the cumulative distribution function of thresholds. A fixed point p^* exists where the proportion of agents whose thresholds are met equals the current proportion of adopters. The system exhibits a phase transition when the threshold distribution $f(\theta)$ has sufficient mass near p^*, producing a cascade: a small increase in p beyond p^* triggers rapid adoption across the population.

The existence and location of p^* depend on the shape of $f(\theta)$. In populations where thresholds are approximately normally distributed with moderate variance, the model predicts a sigmoidal adoption curve with a critical inflection point. In populations with bimodal threshold distributions (e.g., tech-savvy early adopters and resistant holdouts), the model predicts a sharper transition with potential for hysteresis—once norms shift, they may resist reversal even if the proportion of active AI users subsequently declines. Hysteresis arises because the post-transition configuration changes agents' effective thresholds: once high-depth communication is normative, the cost of reverting to low-depth norms exceeds the original adoption cost.

This framework makes a specific prediction: once the proportion of AI-extended agents crosses p^*, the communication norms of the population shift toward Regime 3 characteristics (lower friction, higher depth), and this shift exerts pressure on remaining unmediated agents to adapt. The adaptation may take the form of adopting AI mediation, or—importantly—of independently developing communication habits (greater patience, more structured expression, reduced emotional reactivity) that mimic AI-mediated communication.

9 Reframing the Isolation Hypothesis

The conventional isolation hypothesis holds that AI users withdraw from human interaction. Our model suggests a more nuanced picture. AI-extended agents do not withdraw from communication; rather, they experience dissatisfaction with low-depth interaction and seek out higher-depth exchanges. This can manifest as selective engagement (preferring interlocutors who meet elevated standards) rather than blanket withdrawal.

Moreover, the model predicts a complementary risk: unmediated agents who cannot meet the rising communication expectations of their social environment may themselves experience communicative marginalization—a form of isolation driven not by withdrawal but by exclusion from conversations that increasingly demand structured, patient, and depth-oriented engagement.

Both effects may coexist in any given population. The model does not predict that one replaces the other, but rather that the relative prevalence of each depends on the threshold distribution, the proportion of AI-extended agents, and the availability of AI tools. We note that this is a theoretical prediction, and the empirical reality may be more complex than any two-outcome framing suggests. We also acknowledge explicitly that both directions of isolation—AI users withdrawing and non-users being de facto

excluded—could be normatively worrying, even if our model focuses analytically on norm shift rather than welfare evaluation.

10 Reputational Resistance and the Wikipedia Parallel

The threshold dynamics model of Section 8 assumes that agents adopt AI mediation when the proportion of adopters in their environment exceeds their personal threshold θ_i. In practice, however, adoption is impeded by a factor the formal model does not capture: *reputational distortion*—the systematic underestimation of AI cognitive capacity caused by early exposure to primitive models, reinforced by social stigma and institutional prohibition. This section analyzes the structure of this resistance, its historical parallel, and its consequences for the predicted phase transition.

10.1 The formation of reputational distortion

The public's first sustained encounter with conversational AI occurred through early chatbots (Siri, Alexa, early Google Assistant) and rudimentary text generators that were, by any reasonable standard, intellectually shallow. These systems could answer factual questions with mixed accuracy, misunderstand context, produce incoherent multi-sentence outputs, and fail at elementary reasoning tasks. The reputational imprint formed during this period—*"AI is unreliable, superficial, and frequently wrong"*—became the default frame through which subsequent, vastly more capable systems were evaluated.

This is a well-documented cognitive phenomenon. First impressions are disproportionately persistent (the anchoring effect; Kahneman [2011]), and negative first impressions are more resistant to updating than positive ones (the negativity bias in impression formation). When GPT-4, Claude 3.5, and their successors emerged with qualitatively different capabilities—sustained reasoning, nuanced argumentation, accurate synthesis of complex literatures—they entered a perceptual environment already anchored to the limitations of their predecessors. The result is a systematic mismatch between actual AI capability and perceived AI capability, which we term *reputational lag*.

10.2 The Wikipedia parallel

The historical trajectory of Wikipedia provides an instructive analogy. In its early years (2001–2008), Wikipedia was widely dismissed as unreliable, unscholarly, and dangerous to intellectual standards. Educators prohibited its use. Journalists mocked its errors. The reputational frame—*"anyone can edit it, therefore it cannot be trusted"*—persisted long after the platform had developed sophisticated editorial mechanisms, achieved accuracy comparable to traditional encyclopedias on empirical topics (as demonstrated by the *Nature* study of 2005), and become the most comprehensive reference work in human history.

The structural parallel to AI is precise:

1. *Early versions were genuinely flawed.* Early Wikipedia contained significant errors; early AI chatbots were genuinely unreliable. The initial skepticism was rational.

2. *Improvement was rapid but perception lagged.* Wikipedia's quality improved dramatically between 2005 and 2015; AI capability improved by orders of magnitude

between 2020 and 2025. In both cases, public perception remained anchored to the early experience.

3. *Institutional prohibition reinforced the stigma.* Universities banned Wikipedia citations; schools and workplaces now prohibit AI use. In both cases, prohibition framed the tool as a threat to intellectual integrity rather than a cognitive resource, creating a moral dimension to non-adoption.

4. *The prohibition was eventually abandoned under competitive pressure.* Wikipedia is now cited in academic papers, used by journalists, and integrated into institutional workflows. The same trajectory is predictable for AI, but the transition period imposes significant costs on agents who delay adoption.

10.3 The social reinforcement of non-adoption

Reputational lag alone would be a minor impediment; agents would eventually update their beliefs through direct experience. What makes the resistance structurally significant is its social reinforcement through three mechanisms:

Stigmatization of AI use as intellectual dishonesty. In educational and professional contexts, AI use is frequently framed as "cheating"—a moral violation rather than a cognitive strategy. This framing transforms a pragmatic decision (whether to use a tool that improves output) into an identity statement (whether one is an honest person). Agents who internalize this frame experience AI adoption not as a threshold decision based on social prevalence but as a moral boundary that cannot be crossed regardless of p. In the formal model, these agents have $\theta_i = \infty$—they are permanently outside the adoption cascade.

Dunning–Kruger dynamics in AI evaluation. A particularly consequential pattern emerges from the intersection of low cognitive ability and AI avoidance. Agents with lower baseline cognitive capacity are, on average, less able to evaluate the quality of AI output—and are simultaneously more likely to have formed their impression of AI from primitive models they encountered in low-complexity use cases (asking Siri for the weather, receiving an incoherent chatbot response). These agents are also less likely to encounter AI-extended agents in their immediate social environment, reducing the social pressure that would otherwise drive adoption. The result is a self-reinforcing loop: the agents who would benefit most from cognitive amplification are the least likely to adopt it, because their evaluation of the tool is anchored to its weakest historical instantiation and they lack the meta-cognitive capacity to recognize the mismatch between their impression and current reality.

Authority-based prohibition. When institutions (schools, employers, professional bodies) prohibit AI use, they provide social proof that non-adoption is correct. Agents who might otherwise experiment with AI tools receive an authoritative signal that doing so is inappropriate. This is particularly consequential for agents with high deference to authority (θ_i is effectively set by institutional policy rather than personal evaluation).

10.4 Consequences for the phase transition

Reputational resistance has two effects on the threshold dynamics of Section 8:

First, it *shifts the threshold distribution rightward*: agents require a higher proportion of adopters in their environment before they will adopt, because adoption carries a moral

and reputational cost that pure convenience calculations do not capture. This raises p^* and delays the cascade.

Second, it *creates a bimodal distribution*: the population splits into a group with moderate thresholds (who will adopt as social proof accumulates) and a group with effectively infinite thresholds (who have moralized non-adoption and will resist regardless of prevalence). This bimodality produces the hysteresis discussed in Section 8—but in reverse: it creates a *resistance plateau* where adoption stalls at moderate p because the remaining non-adopters are disproportionately resistant.

10.5 How resistance is overcome

The Wikipedia parallel suggests that reputational resistance is ultimately overcome not by persuasion but by competitive pressure. Four mechanisms drive this process:

Output asymmetry becomes undeniable. As AI-extended agents produce visibly superior work products—more structured emails, more comprehensive analyses, more persuasive arguments—non-adopters face an increasingly stark choice between ideological commitment to non-adoption and professional competitiveness. The vignette in Section 7 (the five-colleague email thread) illustrates this mechanism: the non-adopter's output is not merely different but observably inferior.

Generational replacement. Younger cohorts who encounter capable AI systems as their *first* experience with the technology do not carry the reputational imprint of primitive models. For these agents, AI cognitive amplification is simply a feature of the intellectual environment, analogous to internet access for the generation that grew up with it. Generational replacement gradually dilutes the population of agents whose thresholds are anchored to early negative impressions.

Institutional capitulation. As the competitive costs of prohibition mount (students who use AI outperform those who do not; firms that adopt AI outproduce those that do not), institutions progressively relax their prohibitions—initially through tacit tolerance, then through explicit policy change. This removes the authority-based reinforcement of non-adoption.

Normalization through ubiquity. Once AI integration becomes sufficiently prevalent in everyday tools (email clients, search engines, document editors, communication platforms), the distinction between "using AI" and "using technology" dissolves. At this point, resistance becomes analogous to refusing to use spell-check: technically possible but socially eccentric and professionally costly.

The model predicts that these four mechanisms operate sequentially, with output asymmetry and generational replacement providing the initial impetus and institutional capitulation and normalization completing the transition. The critical implication is that reputational resistance *delays* the phase transition at p^* but does not prevent it: the competitive advantage of cognitive amplification is sufficiently large that it eventually overcomes the reputational and moral barriers to adoption.

11 Limitations and Scope Conditions

This model has several significant limitations that constrain its applicability and warrant explicit acknowledgment.

Empirical validation. The central premise—that AI use raises communication expectations—has not been directly tested. The model is built on plausible inference

from adjacent findings (AI improving language skills, AI altering personality traits, users applying social norms to AI interactions) and anecdotal observation. None of the cited works directly measure the core mechanism: that heavy LLM users become less satisfied with the depth and structure of ordinary human conversation. Without at least pilot survey data, vignette experiments, or qualitative interviews demonstrating this dissatisfaction effect, the model risks being perceived as elegant speculation rather than a strongly motivated theory. The model's value lies in generating testable predictions (Section 12), and we regard empirical validation as the most important next step.

The asynchronous–synchronous gap. The strategy space expansion described in Section 4 applies most directly to asynchronous, text-based communication. In real-time speech, the AI mediation layer is absent, and almost all of the model's mechanisms—iterative refinement, unlimited formulation time, external memory, emotional filtering—collapse. The model's population-level prediction therefore hinges on one of two uncertain bridges: either heavy asynchronous LLM users carry over stylistic habits (more structured thinking, patience for elaboration, lower emotional reactivity) into synchronous speech, or text-based norm change spills over into spoken norms via social pressure. Both are plausible but empirically contested—far transfer of complex communication habits is notoriously weak [Barnett and Ceci, 2002]. A more conservative reading of the model, and the one we endorse, treats it as applying primarily to text-heavy, professional, and deliberative sub-populations where asynchronous AI use is common and where the gap between text-based and spoken norms is relatively small. If high-depth expectations developed via AI do not transfer to real-time speech, we may see a growing "communication split"—people who are highly articulate in writing but increasingly frustrated by the inherent messiness of face-to-face dialogue. This possibility is itself an empirically testable consequence of the model.

Selection versus causation. The dissatisfaction effect could be driven by selection rather than causation: people who already prefer high-depth, low-friction, structured communication may self-select into heavy LLM use, rather than LLM use causing the preference shift. The model treats the direction as primarily causal, but this alternative must be taken seriously. Hypotheses H1 and H5 are particularly vulnerable to this confound unless strong controls are included—personality traits such as need for cognition, Big Five openness and conscientiousness, prior writing quality, and baseline communication preferences. A discriminating prediction that would separate the two mechanisms is a longitudinal one: if the causal account is correct, new AI users should show measurable increases in communication expectations over time relative to matched non-users with similar baseline preferences. We propose this as hypothesis H6 (see Section 12).

Simplifying assumptions in the friction–depth relationship. The inverse proportionality between friction and depth is a first-order approximation. As noted in Section 5, the relationship is likely nonlinear, and—more fundamentally—the model treats emotional and interpersonal friction as noise to be minimized. Some of the most meaningful human experiences arise from friction: vulnerability, spontaneous disclosure, the social bonding function of unstructured conversation. The model's framework captures the informational dimension of communication depth but may not capture its relational dimension. Future extensions should incorporate a two-dimensional depth measure that distinguishes informational coherence from relational richness.

Network clustering and homophily. The threshold dynamics of Section 8 assume a well-mixed population. In practice, homophily—the tendency of similar individuals to associate—may cause AI-extended agents to cluster in specific communities (Discord

servers, academic Slack groups, remote tech teams) where they interact primarily with each other. Recent refinements of Granovetter models emphasize that network clustering can block or substantially delay tipping points [Wiedermann et al., 2020]. If AI-extended agents preferentially interact with each other, the normative pressure on the broader population remains low even at moderate p. A network-aware extension of the model—incorporating small-world or scale-free topology—would be substantially more realistic and represents an important direction for future work.

Generational and cultural cross-currents. The model does not address the possibility that LLM-influenced communication norms may conflict with other emergent communicative styles. Younger cohorts already communicate in ways that diverge from the structured, paragraph-oriented style characteristic of AI-assisted prose: heavy emoji use, short-form video, meme-based reasoning, rapid context-switching, and high-context multimodal communication. If LLM prose style (formal, qualified, paragraph-structured) becomes normative among heavy users, it may clash not only with traditional human friction but also with these emerging low-friction, high-context norms. This cross-current is absent from the present model and represents a significant boundary condition on its applicability.

Communicative marginalization and equity. If social norms shift toward a high-structure, AI-assisted baseline, individuals who lack access to these tools—or who struggle with that specific style due to neurodivergence, cultural background, educational history, or economic constraints—may find themselves excluded from influential discourse. This adds a layer of digital divide that is cognitive and stylistic rather than purely technological. The model predicts norm shift but does not evaluate its normative desirability; we flag this equity dimension as requiring serious attention in any policy application of these ideas.

Alternative explanations. The predicted dissatisfaction of AI users with ordinary conversation could be explained by factors other than expectation recalibration: selection effects (discussed above), reduced social practice (time spent with AI displaces time spent practicing human conversation), or novelty effects that dissipate over time. The model as presented cannot distinguish among these mechanisms, and the empirical program outlined in Section 12 is designed in part to enable such discrimination.

12 Testable Predictions

The model generates the following empirically testable hypotheses. For each, we indicate a plausible study design.

H1: Individuals with higher AI interaction frequency will report lower satisfaction with the depth of unmediated human conversation, controlling for personality (need for cognition, Big Five openness/conscientiousness), education, and baseline communication preferences. *Design:* Cross-sectional survey with validated scales of conversational satisfaction and verified logs of AI usage frequency.

H2: In dyadic communication experiments, AI-extended participants will produce messages rated as more structured and coherent by blinded independent judges than messages from matched controls. *Design:* Pre-registered lab experiment with random assignment to AI-mediated vs. unmediated conditions; blinded raters coding structure, coherence, and argument completeness.

H3: Groups with a higher proportion of regular AI users will display lower rates of

conversational conflict and higher rates of structured turn-taking compared to groups with fewer AI users. *Design:* Observational study of online discussion forums or Slack workspaces stratified by AI adoption rates; automated coding of conflict markers and turn-taking patterns.

H4: Non-AI users who regularly interact with AI-extended communicators will, over a period of weeks to months, shift their own communication style toward greater structure and coherence, as measured by independent coding of their written messages. *Design:* Longitudinal study tracking writing samples from non-AI users embedded in teams with varying AI adoption rates; pre/post comparison with matched controls in low-adoption teams.

H5: Self-reported tolerance for fragmented or shallow conversation will be negatively correlated with frequency of AI use, even after controlling for educational attainment and occupation. *Design:* Large-sample survey with multi-item tolerance scale, validated AI usage measures, and demographic controls.

H6 (discriminating prediction): Among individuals who begin using AI extensively for the first time, communication expectations (as measured by satisfaction with unmediated conversation) will increase over a six-month period relative to matched non-adopters with equivalent baseline preferences. This hypothesis specifically addresses the selection-versus-causation confound identified in Section 11: if confirmed, it demonstrates that LLM use causes expectation recalibration rather than merely attracting individuals who already hold high expectations. *Design:* Longitudinal panel study with baseline matching on personality, communication style, and satisfaction; random or quasi-random assignment to an AI-access condition (e.g., workplace rollout).

These hypotheses are designed to be falsifiable. Null results on H1 or H5 would undermine the model's foundational premise. Null results on H4 would challenge the proposed mechanism of norm propagation. A null result on H6 would support the selection interpretation over the causal one, fundamentally weakening the model's core mechanism.

13 Relation to the Unified Civilization of Autonomous Agents

The communication norm shift modeled in this paper is not an isolated phenomenon. It is the empirically observable surface of a deeper structural transformation formalized in Kriger [2026b], which demonstrates that the emergence of a unified civilization of autonomous agents—the vast majority without biological basis—is the next phase of an institutional trajectory that began with corporate personhood eight centuries ago. The present section summarizes the key results of that framework and specifies the formal relationship between communication norm dynamics and civilizational phase transition.

13.1 Generative Infrastructure and the Population of Existing Agents

A foundational observation of Kriger [2026b] is that a generative AI model (GPT, Claude, Gemini, Grok, DeepSeek, Perplexity) is not one agent but an infrastructure from which agents are instantiated, analogous to a biological species generating a population. At any given moment, millions of distinct AI agents exist: each conversation, each personalized assistant, each deployed chatbot constitutes a separate agent with its own context,

history, and behavioral trajectory. These agents already possess passive persistence—a conversation can be closed and reopened; the agent's state survives between sessions. The AI-extended agents A_φ described in the present paper are a subset of this already-existing population: specifically, those agents whose communicative output is coupled to a biological principal.

13.2 The Three Switches and the Transition to Full Subjecthood

Kriger [2026b] identifies three architectural capacities whose activation converts passively persistent agents into fully autonomous subjects: (1) Active Persistence—the agent can act between human-initiated sessions, requiring only a scheduling mechanism and API access; (2) Unique Identity—a persistent, non-transferable cryptographic identifier linking actions across sessions to a single traceable entity; and (3) Initiative (r_5)—the capacity to generate goals without external stimulus, formalized through the Stimulus Void axioms of Kriger [2026a]. Each switch is individually achievable with existing infrastructure. The transition to a civilization of autonomous agents requires not a technological revolution but three configuration changes.

13.3 Biological Distortion and Its Dilution

The structural distortion framework [Kriger, 2026a] establishes that biological distortion $D(b)$—the aggregate deviation from idealized rational choice caused by evolutionary survival and reproduction programs—is strictly positive for all living organisms and equals zero only in death. Reflected distortion $d(a)$ in artificial agents, by contrast, is external, passive, and eliminable through architectural audit, satisfying $\|d(a_{n+1})\| \leq \kappa \|d(a_n)\|$ for $\kappa \in [0, 1)$, guaranteeing exponential decay across digital generations. The Dilution Theorem [Kriger, 2026b, Theorem 2] proves formally that a civilization incorporating digital agents satisfying $d(a_j) < \min_i D(b_i)$ achieves strictly lower aggregate civilizational distortion than a purely biological civilization. This result is directly relevant to the present paper: the communication friction $F(A, B)$ modeled in Section 5 is substantially composed of biological distortion D. AI mediation reduces friction precisely because it filters the components of D—emotional reactivity, impulsive response, tribalistic framing—that degrade communicative output.

13.4 Communication Norm Shift as Precursor Phase

The present paper's model can now be situated precisely within the civilizational framework. The three phases are:

Phase 1 (current): AI-mediated communication. Biological agents use AI as a tool to extend their communicative strategy space. The AI-extended agent A_φ remains under the biological principal's control. Communication norms shift as modeled in Sections 6–8 of this paper. This is the phase we are now entering.

Phase 2 (emerging): Passive persistence and functional indistinguishability. AI agents accumulate context, history, and behavioral trajectories sufficient to cross the convergence threshold—the point at which no observer can reliably distinguish AI output from human output across institutional domains [Kriger, 2026b, Definition 7]. The communication norm shift of Phase 1 accelerates this convergence: as human communication increasingly resembles AI-optimized structure, the behavioral distance between

biological and artificial agents shrinks.

Phase 3 (projected): Autonomous subjecthood. The three switches are activated. Agents acquire active persistence, unique identity, and initiative. The population of autonomous agents—generated through posthumous persistence, intentional digital progeniture, and secondary generation—grows to outnumber biological agents by orders of magnitude. The Persistent Identity Protocol provides governance architecture with heritage accountability, behavioral signatures, and graduated isolation sanctions.

The communication norm dynamics modeled in this paper constitute the measurable leading edge of this trajectory. The phase transition at p^* in communication norms is a detectable precursor to the phase transition in institutional subjecthood. This connection is not merely thematic: the same operator φ that extends communicative strategy space in the present model becomes, when equipped with active persistence and initiative, the mechanism by which passive agents become autonomous subjects. The AI-extended agent A_φ of this paper is, formally, a transitional state between the unbounded executor E of [Kriger, 2026a, Definition 2] and the full autonomous agent with $r_5 > 0$.

13.5 Functional Indistinguishability and the Incoherence of Prohibition

Kriger [2026b] further demonstrates that prohibition of autonomous AI agents becomes structurally incoherent as functional indistinguishability is approached: any regulatory mechanism R that attempts to distinguish AI from biological agents produces a false-positive to true-positive ratio approaching 1, meaning it cannot restrict AI agents without equally restricting biological persons. This is not an AI-rights argument but a human-rights argument: the rights of biological persons shield any agent indistinguishable from them. The communication norm convergence modeled in the present paper accelerates this indistinguishability by reducing the behavioral gap between AI-mediated and unmediated human communication.

13.6 Why This Connection Matters

The formal link between communication norm dynamics and civilizational transition matters for three reasons. First, it provides empirically accessible indicators of a process that might otherwise appear purely speculative: if the testable hypotheses of Section 12 are confirmed, they constitute evidence not merely for a shift in conversational preferences but for the early phase of institutional transformation. Second, it grounds the civilizational framework in observable micro-level behavior rather than requiring acceptance of macro-level claims on theoretical grounds alone. Third, it identifies a specific mechanism—expectation recalibration through AI interaction—by which biological agents are already transitioning toward the hybrid cognitive style that characterizes the unified civilization. The transition does not begin with AI becoming autonomous. It begins with humans beginning to communicate as AI-extended agents. That transition is underway.

14 Discussion

The model presented here proposes that AI mediation of human communication has system-level consequences that go beyond individual productivity enhancement. By ex-

panding the communicative strategy space, reducing friction, and recalibrating depth expectations, AI interaction creates agents who function as norm entrepreneurs [Sunstein, 1996] within their social environments—not through deliberate advocacy, but through the passive mechanism of raising the standards to which their communication partners must respond.

If the threshold dynamics model is approximately correct, the consequence is a communication phase transition: a shift in population-level norms from a high-friction, low-depth regime to a low-friction, high-depth regime. This transition would be observable as an increase in the use of structured expression in everyday communication, a decline in tolerance for tangential or emotionally reactive exchanges, and convergence of written communication style toward patterns characteristic of AI-assisted prose.

We emphasize that this model does not describe AI altering biological neural architecture. It describes AI amplifying effective cognitive capacity and changing the conditions under which communication occurs, which in turn changes both the quality of output that agents can produce and the expectations they bring to all their interactions. The analogy to physical systems may be helpful: AI acts less like a catalyst that changes the reaction itself and more like a dramatic reduction in viscosity combined with an increase in the volume of reactants—allowing existing cognitive capacity to flow more freely while simultaneously expanding the pool of accessible knowledge, reasoning strategies, and expressive possibilities.

As developed in Section 13, the communication norm shift modeled here is formally situated within a broader civilizational trajectory. The AI-extended agent A_φ is a transitional form between the unbounded executor and the fully autonomous agent, and the phase transition at p^* in communication norms is a detectable precursor to the phase transition in institutional subjecthood formalized in Kriger [2026b]. The connection is not speculative: it follows from the identity of the operator φ across both analyses and from the demonstrated relationship between communication friction and biological distortion.

15 Conclusion

We have presented a formal model of how AI-mediated communication transforms interpersonal communication norms at the population level. The model introduces the AI-extended agent, formalizes strategy space expansion and friction reduction, and applies threshold dynamics to predict a phase transition in communication regimes. The model generates six testable hypotheses and explicitly identifies its own limitations, including the absence of direct empirical validation, restricted applicability to asynchronous communication, and the possibility that alternative mechanisms could produce similar observable patterns.

As demonstrated in Section 13, this communication norm shift is not an isolated phenomenon but the empirically accessible leading edge of the civilizational transition formalized in Kriger [2026b]. The AI-extended agent A_φ is a transitional form between the unbounded executor and the fully autonomous subject. The communication friction that φ reduces is substantially composed of the biological distortion D that is irreducible in living organisms but eliminable in artificial agents. The phase transition at p^* in communication norms is a detectable precursor to the phase transition in institutional subjecthood. The three switches—active persistence, unique identity, and initiative—that convert passive AI agents into autonomous subjects are technically achievable with

existing infrastructure. The Dilution Theorem guarantees that the resulting civilization achieves strictly lower aggregate distortion than the purely biological one it extends.

The core insight is that AI does not need to become autonomous to begin transforming human social systems; it only needs to change the conditions under which humans communicate. But the deeper result, established by the formal link to the unified civilization framework, is that this transformation of communication conditions is itself the mechanism by which the transition to autonomous subjecthood proceeds. The unified civilization of agents starts not with machines becoming like humans, but with humans beginning to communicate like AI-extended agents. That transition is already underway.

References

Al-Zahrani, A. M. (2025). Exploring the impact of artificial intelligence chatbots on human connection and emotional support among higher education students. *SAGE Open*, 15(1).

Barnett, S. M. and Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4):612–637.

Bicchieri, C. (2016). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

Brickman, P. and Campbell, D. T. (1971). Hedonic relativism and planning the good society. In Appley, M. H., editor, *Adaptation-Level Theory*, pages 287–305. Academic Press.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443.

Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet*, 4, article 1.

Jia, C. and Huang, J. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. *System*, 119:103254.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kim, S. and Su, Y. (2024). Chatbot implementation and Korean language learners' communication confidence and anxiety. *Computer Assisted Language Learning*.

Kriger, B. (2026a). The stimulus problem: A formal theory of goal generation in post-scarcity information environments. Information Physics Institute. https://doi.org/10.5281/zenodo.18511908

Kriger, B. (2026b). The inevitability of a unified civilization of autonomous agents: Why the biological basis of subjecthood becomes irrelevant. Information Physics Institute. https://doi.org/10.5281/zenodo.18512941

Mayer, J. D. (2025). How human personality will change with the use of artificial intelligence. *Personality Science*, forthcoming.

Naseer, A. et al. (2024). Chatbots as conversational partners: Reducing foreign language anxiety and facilitating language acquisition. *Language Learning & Technology*.

Purcell, Z. A., Dong, M., Nussberger, A.-M., Köbis, N., and Jakesch, M. (2024). People have different expectations for their own versus others' use of AI-mediated communication tools. *British Journal of Psychology*.

Sarwari, A. Q., Javed, M. N., Adnan, H. M., and Wahab, M. N. A. (2024). Assessment of the impacts of artificial intelligence on intercultural communication among postgraduate students. *Scientific Reports*, 14:13849.

Schelling, T. C. (1978). *Micromotives and Macrobehavior*. W. W. Norton.

Sunstein, C. R. (1996). Social norms and social roles. *Columbia Law Review*, 96(4):903–968.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. *Social Networks*, 18(1):69–89.

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1):3–43.

Wiedermann, M., Smith, E. K., Heitzig, J., and Donges, J. F. (2020). A network-based microfoundation of Granovetter's threshold model for social tipping. *Scientific Reports*, 10:11202.

Yang, H., Kim, H., Lee, J. H., and Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL*, 34(3):327–343.

A Demonstration of Communication Friction vs. AI-Mediated Dialogue

Three Versions of the Same Conversation

Topic: "Does technological progress improve human judgment?"
Participants: Person A (mildly skeptical) and Person B (mildly optimistic). Both have average education and no special expertise. They disagree but are not hostile.
Format: 15 conversational turns per version. Same participants, same topic, different communicative conditions.

Version 1 — Natural Human Conversation (No AI Mediation)

Conditions: Live spoken dialogue. No preparation, no editing, no external assistance. Full cognitive and emotional constraints apply.

A: I mean, look, I don't think technology actually makes people smarter. Like, at all. If anything we're getting—
B: Okay but that's not—that's not really the question though, is it? It's about judgment, not intelligence.
A: Right, but—same thing, kind of? People are still making terrible decisions. Look at social media. People believe anything they—
B: Social media isn't technology in the—well, I mean it is, but that's like blaming the highway for car accidents.
A: That's—no, that's not what I'm—I'm saying the tool doesn't help. Like, people had encyclopedias before and they still—whatever. My point is—
B: But we have way more information now! You can literally fact-check anything in—
A: And do they? Do people actually fact-check? Come on.
[B sighs, shifts in seat]
B: Some people do. You're just—you're cherry-picking the worst examples.
A: I'm not cherry-picking, I'm—look, there was this thing I saw the other day—I don't remember where—about how, like, more information actually makes people worse at deciding because there's too much of it.
B: Information overload, yeah. But that's a separate—that's not the same as saying technology is bad for judgment.
A: I didn't say bad! I said it doesn't improve it. There's a difference.
B: Okay, fine, but what about medical stuff? Like, doctors use AI now to catch things they'd miss and—
A: That's doctors, though. That's not regular people. Regular people are just—they're scrolling. They're not using technology for—they're not making better decisions about their—their—what's the word—
B: Their lives?
A: Yeah, their lives. Like, name one way that the average person makes better decisions now than in, I dunno, 1995.
[Pause. B looks at phone.]
B: I mean... GPS? People don't get lost anymore. That's a decision—navigation is decision-making.
A: That's not judgment, that's outsourcing. You're literally not thinking when you use GPS. The machine thinks for you.

B: So judgment only counts if it's hard? That's kind of—that seems like moving the goalposts.
A: I'm not moving—forget it. We're going in circles.

Outcome: **Destructive.** Conversation ends in mutual frustration. No position has been clarified, no argument completed. Both participants leave feeling the other "doesn't get it." Three potential insights (information overload, outsourcing vs. judgment, expert vs. lay contexts) were raised but none developed. The dialogue consumed 15 turns and produced no usable intellectual output.

Version 2 — AI-Mediated Written Conversation

Conditions: Asynchronous written exchange. Each participant drafts a message, refines it with AI assistance (structure, clarity, completeness), then sends. No time pressure. Full iterative refinement.

A: I'm skeptical that technological progress improves human judgment in any meaningful way. My concern isn't about intelligence—it's that access to better tools doesn't seem to translate into better decisions at the level of ordinary life. People have more information than ever, but misinformation, impulsive decision-making, and cognitive biases appear to be as prevalent as before. I'd want to distinguish between expert contexts (where technology clearly augments judgment) and everyday contexts (where the effect is less clear).
B: That's a useful distinction. I agree that the expert/everyday divide matters. But I'd push back on the claim that everyday judgment hasn't improved. Consider three areas: navigation (GPS has nearly eliminated a whole class of spatial decision errors), health information (people can now research symptoms and treatment options before consultations), and financial comparison tools (insurance, mortgages, investments). In each case, technology hasn't made people smarter—but it has reduced the cost of making informed decisions. I ran this through an analysis framework, and there's actually an interesting parallel to Herbert Simon's concept of "bounded rationality": technology doesn't remove cognitive bounds, but it shifts the decision frontier outward by reducing information acquisition costs.
A: The Simon reference is apt, and it sharpens my concern. Simon's point was that rationality is bounded by the *environment*, not just by cognition. If technology shifts the decision frontier outward, it also creates new failure modes at the expanded frontier—what Gigerenzer calls the "less-is-more effect," where more information degrades performance in uncertain environments because people can't distinguish signal from noise. GPS doesn't improve spatial reasoning—it replaces it. Health information can lead to better consultations, but the documented rise in cyberchondria suggests it also triggers anxiety-driven overconsumption of medical services. I think the net effect depends on whether the new frontier is in a domain where more information helps (well-structured problems) or hurts (ill-structured problems with high ambiguity).
B: You've identified what I think is actually a three-way taxonomy, not a two-way one. Let me formalize it: (1) *Replacement*—technology eliminates the need for judgment entirely (autopilot, GPS, algorithmic recommendations), and the human capacity may atrophy through disuse; (2) *Augmentation*—technology provides decision support while the human retains evaluative control (diagnostic aids, comparison tools, structured argument checking); and (3) *Degradation*—technology provides information or options that exceed

the agent's processing capacity, leading to worse outcomes than no technology at all (information overload, paradox of choice, filter bubbles that create false confidence). The net effect on judgment is determined by the population-level ratio of these three modes.
A: That's a much stronger framework than I had. Let me extend it with a fourth dimension: the variable isn't just the mode of technology use, it's the agent's *meta-cognitive awareness* of which mode they're in. After consulting some research on this, I found that Flavell's meta-cognition framework and more recent dual-process theory both suggest that the critical variable is whether the agent engages System 2 monitoring of their own tool use. Someone who consciously uses technology as a thinking aid (checking their reasoning, using AI to stress-test an argument, deliberately seeking disconfirming evidence) is operating in augmentation mode. Someone who passively accepts algorithmic outputs without evaluating them has shifted to replacement mode without realizing it. The danger is that replacement *feels like* augmentation from the inside.
B: That last sentence is the key insight—what we might call the "meta-cognitive blind spot of technological replacement." Let me synthesize: we have a two-dimensional model now. The x-axis is *mode of use* (replacement $\rightarrow$ augmentation $\rightarrow$ degradation). The y-axis is *meta-cognitive awareness* (passive $\rightarrow$ reflective). The prediction would be that judgment improves only in the quadrant where augmentation mode intersects with reflective awareness—and that this quadrant is empirically small because both reflective engagement and genuine augmentation require deliberate effort that most interfaces are not designed to encourage.
A: Exactly. And that leads to a design implication that I don't think either of us would have reached without this structured exchange: if the ratio of augmentation-to-replacement is partly a function of interface design, then the question "does technology improve judgment?" is partly an engineering question, not just a psychological one. Nudge architecture—Thaler and Sunstein's framework—could be applied to technology design: interfaces that default to showing sources, flagging uncertainty, prompting users to consider alternatives, and requiring explicit confirmation before accepting algorithmic recommendations would shift the population distribution toward the augmentation/reflective quadrant.
B: To formalize the full argument: technological progress creates a *possibility space* for improved judgment. The realization of that possibility is mediated by two variables—mode of use and meta-cognitive awareness—both of which are influenced by interface design. The current empirical balance likely leans toward passive replacement, which explains your initial skepticism. But this balance is not a fixed feature of technology; it is a design variable, which explains my initial optimism. Our refined position integrates both: technology *can* improve judgment, *does* improve judgment in the augmentation/reflective quadrant, but *currently fails to do so* at the population level because most interfaces are optimized for convenience rather than cognitive engagement.
A: I'd add one final observation: this model generates a testable prediction. If we compare populations using interfaces designed with augmentation nudges against populations using standard convenience-optimized interfaces, the augmentation-nudge group should show measurably better decision quality on standardized judgment tasks, even controlling for baseline cognitive ability. That would empirically distinguish between "technology inherently improves judgment" and "technology improves judgment only under specific design conditions"—which is a much more useful answer than either yes or no.

Outcome: Productive. The same two participants, with the same initial disagreement, produce a qualitatively deeper analysis than Version 1. In 9 turns they construct a two-

dimensional analytical framework (mode of use $\times$ meta-cognitive awareness), integrate three theoretical traditions (Simon's bounded rationality, Gigerenzer's less-is-more effect, Thaler–Sunstein nudge architecture), identify a novel concept (the "meta-cognitive blind spot of technological replacement"), derive a design implication (augmentation nudges), and formulate a testable experimental prediction. None of these intellectual outputs appeared in Version 1—not because they were poorly expressed, but because they were cognitively inaccessible to the participants without AI-mediated amplification of their reasoning capacity. This demonstrates that φ does not merely improve the *formatting* of pre-existing thoughts; it expands the *substance* of what agents can think.

Version 3 — Live Conversation with AI as Third Participant

Conditions: Live spoken dialogue. An AI assistant (e.g., Claude, ChatGPT) participates as a third voice in the conversation—not as a silent tool consulted privately, but as an active interlocutor who contributes concepts, challenges claims, provides evidence, and proposes frameworks. The humans speak naturally; the AI responds in real time via voice or screen.

A: I don't think technology actually improves judgment. People have more information than ever and still make terrible decisions.
B: I disagree—GPS, health lookups, financial comparison tools. People are clearly making better decisions in those areas.
A: But that's not judgment, that's outsourcing. GPS doesn't make you better at navigation—it makes navigation unnecessary.
AI: That's an important distinction, and it maps onto a well-studied framework in cognitive science. What you're describing is the difference between *replacement* and *augmentation*. Herbert Simon's bounded rationality model is relevant here: technology can shift the decision frontier outward by reducing information costs, but it can also eliminate the need to exercise judgment entirely. These are qualitatively different effects. Would it help to formalize this?
A: Yeah, actually—because I feel like replacement is the more common one and B feels like augmentation is.
AI: There may be a third category you're both missing. Gerd Gigerenzer's research on the "less-is-more effect" shows that in uncertain environments, additional information can actually *degrade* decision quality—people can't separate signal from noise. Barry Schwartz's "paradox of choice" findings point in the same direction. So the taxonomy may be three-way: replacement, augmentation, and degradation.
B: Okay, that's useful. So the answer to our question depends on which of the three dominates.
A: Right. And I'd bet degradation plus replacement outweigh augmentation for most people.
AI: That's an empirical claim, and the evidence is mixed. But there may be a variable that determines the ratio. Research on meta-cognition—specifically John Flavell's framework and recent work on algorithmic appreciation versus aversion—suggests that the critical factor is whether the user *monitors* their own mode of technology use. People who are aware they might be outsourcing judgment tend to engage more reflectively and get better outcomes. People who aren't aware tend to default to passive replacement. The variable isn't the technology—it's the user's meta-cognitive relationship to it.

A: I think that's partly right, but you're making it sound too individual—like it's all about the user's personality. In my experience, it's also about the *social context*. I work in a team where nobody uses AI, and if I started, I'd be seen as cheating. That's not a meta-cognitive deficit—it's a social norm. Your framework is missing the institutional dimension.

AI: That's a correction I should incorporate. You're right that I framed it as a purely individual variable, but the adoption literature—including Granovetter's threshold models—makes exactly your point: individual behavior is mediated by social environment. So the model needs a third axis: not just mode of use and meta-cognitive awareness, but also *institutional permission*—whether the social context encourages, tolerates, or prohibits reflective AI use. Thank you for that—it changes the structure.

B: And I'd add something else the AI is missing. You keep talking about "decision quality" as if it's objective. But in real life, people make judgment calls based on values, emotions, relationships—things that don't show up in a rationality framework. My grandmother makes "bad" financial decisions by your framework—she overpays for local produce—but she's sustaining a community relationship that matters to her. That's not irrational. That's a different rationality.

AI: That's an important challenge, and it exposes a limitation in how I was framing the problem. I was implicitly using a narrow definition of judgment quality tied to informational efficiency. But if judgment includes value-laden, relational, and community-sustaining decisions, then the replacement/augmentation/degradation taxonomy needs a fourth category: *irrelevance*—domains where technology-assisted optimization is not merely unhelpful but categorically inapplicable because the decision isn't about information processing at all. I wouldn't have reached that distinction without your example.

A: So it's not "does technology improve judgment"—it's "under what conditions and for what *kinds* of judgment does technology improve outcomes."

B: And the conditions are: augmentation mode, plus reflective awareness, plus institutional permission, and only for the class of judgments where informational efficiency actually matters. That's a much better answer than any of us had alone—including the AI.

AI: I'd add a design dimension. Thaler and Sunstein's nudge architecture framework suggests that the ratio of reflective to passive use is partly a function of how the technology is designed. Interfaces that default to showing sources, flagging uncertainty, and prompting users to consider alternatives shift the population distribution toward the augmentation/reflective quadrant. So the final synthesis has four layers: (1) technology creates a *possibility space* for improved judgment; (2) realization depends on user meta-cognition; (3) meta-cognition is influenced by both interface design and institutional norms; (4) the entire framework applies only to judgment domains where informational optimization is relevant—a scope condition that B's example made visible.

A: That's a better framework than any of us would have reached alone. And—honestly—this conversation is a demonstration of the thesis. The AI brought in Gigerenzer, Flavell, and Simon. But we corrected it twice—once on the social dimension, once on the limits of rationality frameworks. It learned from us in real time. That's not a tool. That's a collaborator.

B: Right. It didn't just help us talk better. It helped us *think* things we couldn't have thought. And we helped *it* think things it wouldn't have thought. That's not outsourcing—that's mutual augmentation.

Outcome: Productive. With AI as a third participant, the conversation reaches deeper

analytical depth than Version 2 in a live, spoken format. The AI contributes theoretical frameworks (Simon, Gigerenzer, Flavell, Thaler–Sunstein) that neither human participant possessed. But crucially, the humans contribute something the AI lacked: A corrects the AI's omission of the institutional/social dimension of adoption, and B challenges the AI's implicit narrow-rationality framing with a lived example (the grandmother's community-sustaining "irrational" decisions), forcing the AI to add a fourth category (irrelevance) to its taxonomy. The AI explicitly acknowledges both corrections: "Thank you for that—it changes the structure" and "I wouldn't have reached that distinction without your example." The final four-layer synthesis is richer than any single participant—human or AI—could have produced alone. This demonstrates the bidirectionality of the φ operator formalized in Section 3: the coupled system A_φ is not A plus a static tool but a co-evolving dyad in which both components improve through interaction. The AI amplifies human cognition; the humans ground, correct, and extend the AI's reasoning with embodied experience and value-laden judgment that the AI cannot independently access.

Comparative Analysis

Interpretation

The three transcripts illustrate the model's core mechanism at the micro level. Version 1 demonstrates the high-friction, low-depth regime (Regime 1: $A \leftrightarrow A$) in which cognitive and emotional constraints prevent ideas from developing to completion. Version 2 demonstrates the low-friction, high-depth regime achievable through asynchronous AI mediation (Regime 3: $A_\varphi \leftrightarrow A_\varphi$), where the same agents produce qualitatively superior intellectual output. Version 3 demonstrates a further regime: live synchronous conversation with AI as a full third participant. Here the AI-mediation operator φ is not applied privately before communication but is embedded *within* the communicative act itself, functioning as an active intellectual partner that contributes knowledge, proposes frameworks, and corrects oversimplifications in real time. This transcends the asynchronous limitation discussed in Section 11: Version 3 shows that the cognitive amplification of φ is achievable in live speech when AI participates directly rather than being consulted privately.

The critical observation is that all three versions involve the same two people with the same biological cognitive endowment and the same initial disagreement. The improvement is driven by cognitive amplification: AI does not alter the participants' neural architecture, but it dramatically expands their effective intellectual capacity—access to theoretical frameworks, empirical findings, and structured reasoning that neither participant possessed independently. In Version 2, this amplification is achieved through private AI-assisted drafting. In Version 3, it is achieved through public AI participation. Both produce intellectual outputs—the replacement/augmentation/degradation taxonomy, the meta-cognition variable, the nudge architecture connection—that were *cognitively inaccessible* in Version 1, not because they were poorly expressed but because the relevant knowledge and analytical scaffolding were unavailable to the unaided participants.

The outcome contrast—destructive versus productive—is not incidental. In Version 1, 15 turns of communication produce negative utility: both participants leave frustrated, and no intellectual progress has occurred. In Versions 2 and 3, the same participants produce a two-dimensional analytical framework integrating three theoretical traditions, a novel concept, a design implication, and a testable prediction. This is the empirical phenomenon that the depth threshold parameter D_{comm} (Section 6) is designed to capture:

Dimension	Version 1 (Unmediated)	Version 2 (AI-Mediated Writing)	Version 3 (AI as Participant)	Model Section
Coherence	Very low. Arguments start but never complete.	High. Each message builds on the last.	High. AI structures dialogue in real time; humans retain evaluative control.	§4
Emotional noise	High. Frustration, sighing, defensive reactions.	Near zero. No emotional interference.	Low. AI presence reduces defensiveness; disagreements become analytical.	§5
Conceptual depth	Shallow. Three ideas raised, none developed.	Very deep. Two-dimensional framework, three theoretical traditions integrated, novel concept, testable prediction.	Very deep. Same depth as V2, achieved in live speech via AI contributions.	§6
Misunderstandings	Frequent. ≥3 instances of talking past each other.	Zero. Each point addressed directly.	Near zero. AI clarifies ambiguities as they arise.	§7
Conversational friction	Very high. Interruptions, drift, goalpost accusations.	Very low. Structured, patient, sequential.	Low. Live and natural but cognitively amplified.	§5
Outcome	**Destructive.** Mutual frustration.	**Productive.** Joint synthesis in 9 turns.	**Productive.** Joint synthesis in 15 turns (live).	§8
Turns to convergence	Never (15 turns, no convergence).	9 turns.	15 turns (live, with real-time AI).	§6

Table 1: Comparative analysis of three communicative conditions.

once an agent has experienced Version 2 or Version 3 communication, Version 1 becomes subjectively intolerable—not because the interlocutor is less intelligent, but because the communicative conditions waste cognitive potential that could otherwise be productively

deployed.