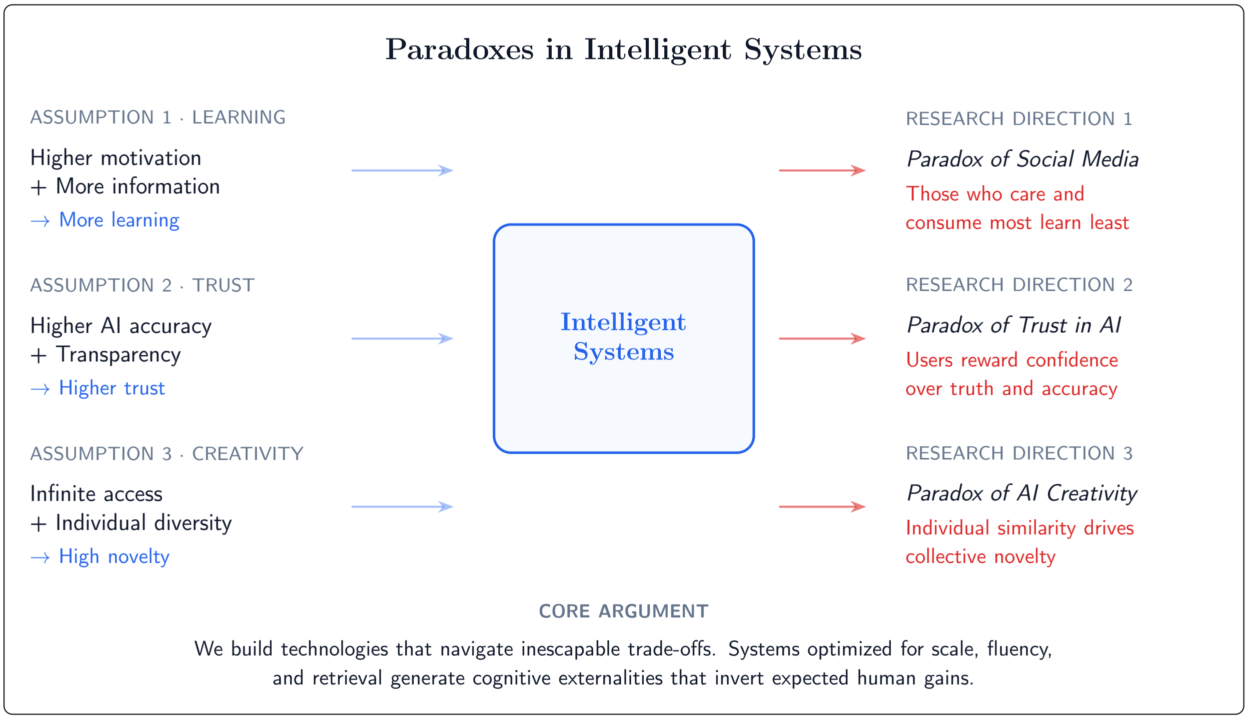
Paradoxes in the interaction between intelligent systems and human cognition

Current scholarship often characterizes AI as an additive tool: a means to provide more information and faster analysis. This perspective posits a linear relationship in which human capacity scales with model capability, suggesting that increased model accuracy and information abundance should yield smarter and more discerning individuals. However, empirical reality contradicts this view. Society exists in a landscape of unprecedented information abundance, yet learning has not increased commensurately. Models are increasingly accurate, yet trust in these systems remains stagnant. Generative AI produces immense volumes of content, yet often fails to yield genuine novelty under scrutiny.

My work spans three directions. Each isolates a paradox created by intelligent systems in which expected cognitive gains are inverted. These paradoxes show that human learning, trust, and creative discovery behave differently once AI mediates the flow of information. I identify a set of puzzling cognitive externalities that remain largely unexplained within computer science, political science, and psychology, fields that have traditionally treated information, reliability, and diversity as stable resources. While prevailing theories assume that AI mediated interaction should improve learning and innovation in a linear manner, I demonstrate that the architecture of these systems often produces the opposite result. I treat paradoxes in learning, trust, and creativity as early signals of a broader question about the trajectory of human progress by redefining who controls access to truth, who is able to learn, and whose ideas are allowed to grow.



We build technologies that navigate inescapable trade-offs. Systems optimized for scale, fluency, and retrieval generate cognitive externalities that invert expected human gains.

Figure 1: *My research roadmap: how conventional assumptions lead to paradoxical outcomes through intelligent systems.*

1. Algorithmic Architecture and Learning

In this line of research, I examine how the design of social media affects political learning. While classical theories posit that motivation is the primary gatekeeper of knowledge, my work shows that algorithmic attention allocation has inverted this relationship. I investigate this inversion by asking how the information environment forces contact with politics for some, while saturating others with more than they can process.

I built an empirical methodology that links fine-grained digital trace data, logged over twelve consecutive months, to repeated surveys of the same individuals across an eighteen-month panel. Using a daily-level measure of incidental exposure, I find that individuals who rarely seek politics still gain structure in their attitudes. As their browsing environment introduces content through pathways they did not select, their views align more coherently with a single ideological direction. This suggests that disengagement no longer insulates people from influence. Instead, political information unintendedly enters through the back door via entertainment and social browsing.

Furthermore, I examine how this same system impacts highly engaged users. I built measures of information overload quantifying volume, diversity, and session fragmentation, which I paired with repeated knowledge assessments across eight survey waves. I find a clear ceiling: as exposure grows, knowledge initially rises but then plateaus, and in some cases declines. High-interest individuals expend attention to stay informed, yet the system outpaces them. My findings indicate that AI-driven delivery spreads cognitive resources too thin, meaning more news does not translate into more uptake once the environment surpasses human processing capacity. My work is the first in political science to formally study the effect of information overload on the political attitudes of citizens.

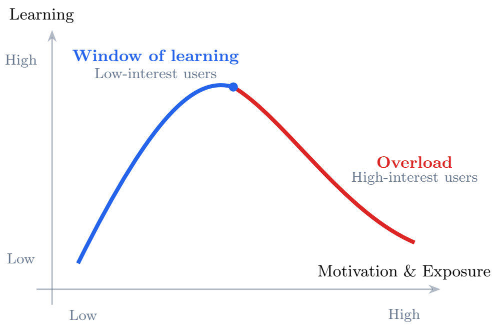


Figure 2: Learning increases with motivation and exposure up to a ceiling, after which overload reduces uptake.

My results make the paradox visible: In a pre-social media era, scarcity made motivation the primary factor in learning. In AI-mediated environments, forced exposure and processing overload are the determinant forces. The disengaged are nudged into politics, while the engaged encounter volumes they cannot convert into competence. This work replaces outdated input-output models with a framework grounded in cognitive limits, treating the environment, not just the individual, as a central driver of who learns and who falls behind.

The Paradox of Social Media and Learning

We built social media to democratize information and make learning effortless. My findings show the result is the opposite. The very users who care most are swamped with more than they can process and learn less. Meanwhile the least interested learn more because the platform architecture pushes political content into their feeds whether they seek it or not. A system designed to expand knowledge ends up rewarding disengagement and penalizing motivation.

2. Trust, confidence, and human evaluation of AI systems

My second line of research examines the *confidence calibration paradox.* Existing scholarship treats AI overconfidence as a design flaw, and argues we should build systems that openly admit uncertainty. Yet, a long tradition in social psychology shows that humans reward overconfidence, regardless of truth. I demonstrate that these perspectives collide once people evaluate AI systems they cannot verify: **AI confidence becomes a trade-off between epistemic honesty and practical persuasion**. I show that when models disclose uncertainty they become more trustworthy in principle, but less trusted in practice, because users penalize the very honesty experts demand. This inversion explains why well intentioned reforms toward honesty can lower trust since the social incentive structure rewards overconfidence and penalizes transparency. My research is the first to treat AI confidence as a relationship between what the model actually believes and how strongly it presents that belief, and re-frames the core problem of trustworthy AI in which making AI honest creates a trade off with its practical utility.

To isolate the mechanisms of this failure, I led a team of researchers at FAR.AI, Harvard, McGill, and Mila to build a fact-checking platform that logs user interactions with large language models. In a first preregistered experiment, I independently varied the linguistic confidence of the model while keeping factual content constant. I find that people rate assertive models as more accurate and credible, confirming that the surface tone of an answer heavily shape trust and perusasion. Users rewarded linguistic confidence even when the model lacked internal evidence, establishing that persuasion often overrides accuracy in opaque environments.

Furthermore, I examine how transparency alters this dynamic. In a second preregistered experiment, I randomly assign to respondents the display of the model's internal certainty via quantitative belief scores derived from token probabilities. I find that when users can see the numeric score, they penalize rhetorical confidence when it exceeds the internal belief. The mismatch between internal and external (linguistic) certainty becomes a visible violation of the system's epistemic honesty, causing users to withdraw trust.

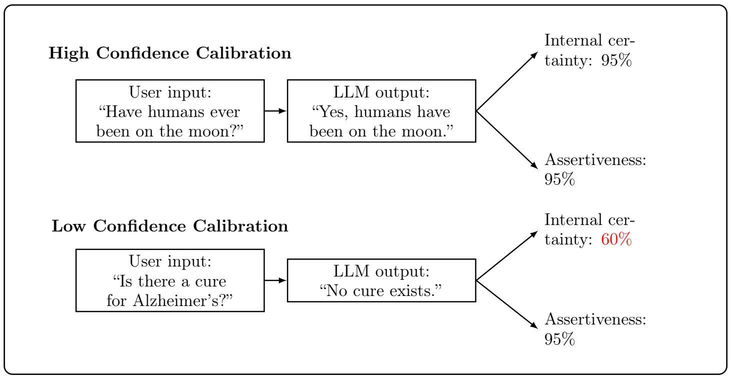

High Confidence Calibration
User input: "Have humans ever been on the moon?"
LLM output: "Yes, humans have been on the moon."
Internal certainty: 95%
Assertiveness: 95%
Low Confidence Calibration
User input: "Is there a cure for Alzheimer's?"
LLM output: "No cure exists."
Internal certainty: 60%
Assertiveness: 95%

Figure 3: Two examples with differing calibration. The upper one is well calibrated. The lower one is overconfident given low internal certainty.

In summary, I show that confidence in AI should be understood as the alignment between what a model believes and how confidently it speaks. I am the first to demonstrate that systems rewarded for assertiveness in opaque settings are punished for that same behavior once their uncertainty is visible. My findings establish that trustworthy AI requires calibration across both internal belief and external expression because, without structural safeguards, the social environment selects for models that sound convincing rather than models that actually know when they are wrong.

The Paradox of Trust in AI

We build AI systems to be accurate and honest so people can trust them. My findings show that the result is inverted. When models admit uncertainty, users trust them less even when they are correct. When models express unwarranted confidence, users trust them more even when they are wrong. A system designed to encourage transparency ends up rewarding overconfidence and punishing honesty.

3. AI and the structure of collective knowledge

My final line of research investigates how generative AI reorganizes the structure of collective knowledge. Current scholarship predominantly characterizes AI-driven homogenization as a normative negative, arguing that by compressing variance, these systems erase diversity and suppress creativity. I challenge this consensus by developing a formal and empirical account of the *Variance–Novelty paradox*. While I acknowledge that AI flattens informational diversity within domains, I demonstrate that this very mechanism paradoxically increases global novelty by lowering translation costs and allowing for new connections across previously disconnected fields.

I begin by formalizing the mechanism of homogenization through what I term the *AI Prism*. I build a theory linking reduced lexical diversity to *AI derivative epistemology*, a mode of inquiry where humans outsource evidentiary work to statistical models. I show that the pairing of model loss-reduction objectives with user deference systematically compresses informational variance. When foundation models mediate communication, stylistic and conceptual distributions converge, transforming variance, traditionally viewed as epistemic capital, into homogenized output.

However, I confront the prevailing view that this variance reduction inevitably stifles discovery. Contrary to this perspective, I argue that local variance compression is the precise mechanism that unlocks global innovation. As generative models standardize specialized languages, they transform heterogeneous representations into interoperable modules, effectively "liquefying" meaning. Domain-specific jargon is translated into a shared statistical vocabulary, lowering the transaction costs of moving ideas across institutional and disciplinary boundaries.

To test this dynamic, I build agent-based models simulating how AI adoption shapes knowledge over time. I find a characteristic U-shaped trajectory: initially, compressed representations facilitate novel pairings between previously unconnected clusters, increasing global novelty. However, as epistemic deference increases and synthetic data overwhelms human input, the system transitions toward informational monoculture. In summary, my work inverts the standard critique of homogenization. I demonstrate that the mechanism flattening outlier expressions is the same one that generates structurally new ideas. Paradoxically, local homogenization becomes a precondition for global novelty.

The Paradox of AI-Driven Innovation

We built generative AI to unlock the full scope of human knowledge and spark new ideas. My findings show that the outcome flips our expectation. These systems compress expression into familiar and repeated patterns that limit individual originality, yet this very standardization lowers translation barriers and enables ideas to travel farther. A technology designed to retrieve anything can barely produce novelty on its own, but by reducing variance it paradoxically increases cross-domain recombination and expands collective discovery.

Across all of my projects, my results point to one core conclusion: We build technologies that force trade-offs. For social media, we designed systems that push so much information that people feel overwhelmed. Yet those who care the least are learning more. This reveals a trade-off over who in society ends up gaining knowledge. For AI transparency and confidence, my findings show a trade-off between honesty and practical utility. We can make AI more honest but that makes it a worse tool. Any move that favors one objective harms the other. For AI driven creativity and innovation, my findings show that AI generated content will make many ideas look similar and lose uniqueness because AI will mediate most knowledge production. Yet this reduction in variance will increase the chance that humans borrow from different domains, learn from different domains, and combine ideas across domains without any communication barrier. That change will lead to more innovation and creativity at the collective level. Each intelligent system achieves one goal only by undermining another.

4. Future Work: Mass Delegation to AI Agents

My completed projects establish that AI and digital infrastructures intervene directly in human cognition. My future work extends this foundation by developing and testing a predictive science of mass delegation to AI systems. I contend that delegation to AI agents will become the dominant mode of human–machine interaction in the coming decade, yet the field lacks the conceptual and empirical tools necessary to anticipate its consequences.

Human-agency expansion has been treated as an unquestioned virtue in the design and evaluation of AI. The assumption has been that more control, more choice, and more cognitive leverage automatically translate into human flourishing. Yet, AI breaks that logic. Increasing agency now enables new forms of over-reliance, strategic abdication, and harmful cognitive offloading. Conversely, decreasing agency can produce positive outcomes when automation protects users from tasks that exceed human processing limits or from environments designed to exploit human weaknesses. We currently have no account that distinguishes when human agency expansion is beneficial and when it is corrosive, nor when agency reduction becomes a protective intervention rather than a threat. My future work targets this blind spot by building the frameworks and empirical evidence required to understand the effects of mass delegation of to AI on human agency.

I will lead a research program that explains when delegation expands human agency and when it erodes it. Rather than assuming that more assistance is always beneficial, my work will identify the conditions under which AI support produces safety, and the conditions under which it produces dependence. I will build the first framework capable of predicting how design choices in assistance systems scale from individual decision making to societal outcomes.

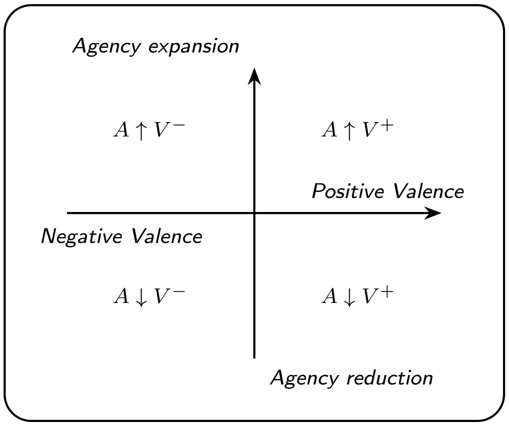


Figure 4: Outcome regions for AI delegation effects.

I plan to structure this inquiry along three dimensions: (1) A *moral dimension* that traces whether delegated decisions align with expressed human intentions; (2) A *temporal dimension* that tracks long-term changes in skill acquisition versus atrophy; (3) A *collective dimension* that models how many small acts of delegation aggregate into shifts in social capacity, civic participation, and autonomy.

Methodologically, I will execute this agenda through a coupled experimental–computational approach. I will design behavioral experiments that quantify how often users allow AI systems to override their stated preferences, and how different assistance philosophies — such as direct recommendations versus metacognitive nudges — affect long-term human competence. I will then embed those empirically estimated parameters into agent-based simulations to evaluate how different delegation regimes scale, identifying early-warning indicators of harmful dependency.

In sum, I aim to build the theoretical and computational tools required for a predictive science of AI delegation. Effective governance of intelligent systems requires more than measuring model accuracy. It requires understanding how widespread reliance reshapes human intent, capability, and democratic agency. My long term goal is to ensure that AI systems are built and governed in ways that strengthen the cognitive foundations of human self-determination.