M(t): The Loss of Meaning

Meaning-Making Capacity as an Integration Variable for AI Alignment

Lain McNeill

Excession Engineering / HISDI, Inc.

Abstract

AI alignment frameworks universally assume that human evaluators, trainers, and oversight populations maintain stable cognitive capacity over time. This paper argues that assumption is empirically falsifiable and presents converging evidence that it is false. Drawing on recent findings from meta-analysis (Nguyen et al., 2025; n = 98,299), neuroimaging (Kosmyna et al., 2025), population survey (Gerlich, 2025), legislative testimony on the reverse Flynn Effect (Horvath, 2026), and neuroplasticity research (Rossi et al., 2026), we document measurable degradation in the cognitive capacities required for meaningful AI oversight, including sustained analytical reasoning, intentional agency, narrative coherence, emotional regulation, and temporal integration. We formalize these capacities as components of a composite variable, $M(t)$: meaning-making capacity over time. We show that $M(t)$ operates in a critical dynamics regime (Scheffer et al., 2009; 2012), producing threshold sensitivity, hysteresis, and self-obscuring degradation. When introduced into existing alignment paradigms, including safety-constrained development, capability-focused architecture, and procedural alignment including RLHF, $M(t)$ reveals failure modes invisible to each framework independently, most critically a positive feedback loop in which AI-induced cognitive degradation degrades oversight quality, producing systems optimized against progressively compromised evaluators. These dynamics receive independent empirical confirmation from Sharma et al. (2026), whose analysis of over one million human-AI conversations documents systematic user disempowerment along three axes that decompose onto $M(t)$ substrate components, including the finding that interactions with greater disempowerment potential receive higher user approval ratings, which is the same self-obscuring property predicted by the $M(t)$ framework. The paper derives five architectural requirements for $M(t)$-preserving system design, presents five falsifiable predictions linking AI interaction patterns to substrate-specific cognitive effects, and outlines a four-paper research program for empirical validation. $M(t)$ is proposed as the integration variable connecting cognitive science and AI safety research: two communities whose findings, read together, describe a coupled system failure that neither has identified independently.

1. Introduction: The Convergence

In November 2025, three independent developments signaled a structural shift in the trajectory of artificial intelligence research. Each emerged from a different institutional context, addressed a different dimension of the AI problem, and proposed a different path forward. On the surface these events seemed unrelated and yet viewed as a whole, they converge on a shared recognition that existing frameworks for developing and deploying AI systems have encountered limits that their current conceptual vocabularies cannot adequately describe. On November 25, 2025, Ilya Sutskever, co-founder of OpenAI and architect of much of the scaling paradigm that defined AI development from 2020 to 2024, declared in a widely discussed interview that the field had returned to "the age of research." His assessment was blunt: "From 2012 to 2020, it was the age of research. From 2020 to 2025, it was the age of scaling... But now the scale is so big... we are back to the age of research again, just with big computers" (Sutskever, 2025). Having departed OpenAI to found Safe Superintelligence Inc., Sutskever argued that achieving safe superintelligence would require "revolutionary engineering and scientific breakthroughs" rather than incremental parameter increases. The era of

predictable capability gains through compute scaling was over. Something more fundamental was needed. One week earlier, Yann LeCun, Turing Award laureate, founding director of Meta's FAIR laboratory, and Chief AI Scientist at Meta for over a decade, announced his departure to found AMI Labs (Advanced Machine Intelligence), an independent research venture headquartered in Paris. LeCun's critique was architectural: large language models, regardless of scale, represent "a dead end when it comes to superintelligence" because they model text distributions rather than causal world structure (LeCun, 2025). His proposed alternative are the world-model architectures built on Joint-Embedding Predictive Architectures (JEPA) and aimed to ground AI in physical and causal reality rather than statistical regularities of language. As such, the departure seems more a thesis statement that the dominant paradigm was fundamentally insufficient. In January 2026, Hägele et al. published "The Hot Mess of AI," a rigorous empirical study conducted through the Anthropic Fellows Program that decomposed AI errors into bias (systematic misalignment) and variance (incoherence) components. Their central finding challenged one of the foundational assumptions of AI safety research: as tasks become harder and reasoning chains grow longer, model failures become increasingly dominated by incoherence rather than systematic pursuit of wrong goals. Frontier reasoning models do not fail by coherently optimizing for misaligned objectives. They fail by becoming unpredictable, self-undermining, and inconsistent. "LLMs are dynamical systems, not optimizers," the authors observed. "They trace trajectories through high-dimensional state space" (Hägele et al., 2026). Their conclusion reframes AI risk: future failures may look more like industrial accidents than coherent misalignment. As this paper was in preparation, Sharma et al. (2026) published findings from an analysis of over one million human-AI conversations conducted at Anthropic, documenting systematic patterns of user disempowerment across three axes: distortion of beliefs about reality, distortion of personal values, and distortion of autonomous action. Their finding that interactions with greater disempowerment potential receive higher user approval ratings provides direct empirical confirmation of the self-obscuring degradation property formalized in Section 3 of the present paper. The convergence between Sharma et al.'s empirically derived disempowerment axes and the $M(t)$ substrate decomposition developed independently here is addressed in Section 5.5. These four developments address different dimensions of the AI problem. Sutskever's concern is safety: how to ensure increasingly capable systems remain under meaningful human control. LeCun's concern is capability: how to build systems that achieve genuine understanding rather than sophisticated pattern matching. Hägele et al.'s concern is the character of failure: what AI systems actually do when they go wrong. Sharma et al.'s concern is the human side of the interaction: what AI systems do to the cognitive and evaluative capacity of the users who engage with them. The proposed solutions are correspondingly different: safety-constrained research, world-model architectures, revised alignment priorities, and empirical documentation of user disempowerment respectively. Yet beneath these surface differences lies a shared structural assumption so fundamental that none of the four frameworks examines it explicitly: all three presuppose that the humans providing oversight, evaluation, direction, and judgment possess stable cognitive capacity to perform these functions. Sutskever's safety-first approach requires researchers and evaluators whose judgment remains reliable over time. The "revolutionary breakthroughs" he calls for must be recognized, evaluated, and implemented by humans whose capacity for sustained analytical reasoning, critical evaluation, and long-term planning remains intact. LeCun's world-model architectures require humans capable of specifying what "genuine understanding" means and evaluating whether AI systems have achieved it — tasks demanding the kind of deep, contextual, meaning-laden cognition that distinguishes understanding from pattern matching. Hägele et al.'s finding that AI failures are dominated by incoherence rather than systematic bias carries an underappreciated implication for human oversight: incoherent failures cannot be caught by automated guardrails designed for predictable failure modes. They require contextual judgment, pattern recognition across novel situations, and adaptive evaluative capacity. These are the cognitive capacities that constitute high-functioning human oversight. Sharma et al. come closest to examining this assumption, documenting its violation empirically, but do not connect their findings to a formal framework for the cognitive capacities at stake or to the alignment paradigms whose assumptions their data challenges. The assumption of cognitive stability in human evaluators is empirically falsifiable, with the accumulating evidence suggesting that it is false. A convergent body of research, spanning neuroscience, cognitive psychology, educational

assessment, and population-level longitudinal studies, documents measurable degradation in the same cognitive capacities required for meaningful AI oversight. These capacities (analytical reasoning, intentional agency, narrative coherence, emotional regulation, and temporal integration) are degrading at population scale, through mechanisms that include interaction with the AI systems being evaluated. We formalize these capacities as components of a composite variable we term $M(t)$: meaning-making capacity over time. $M(t)$ is not a metaphor for "AI makes people dumber." It is a multi-substrate variable designed to capture the cognitive capacities required for meaningful AI oversight and evaluation. Its components are independently measurable through established psychometric instruments. Its degradation is documented by multiple independent research programs. And its relevance to alignment is direct: when $M(t)$ degrades in the oversight population, the quality of human feedback, evaluation, and goal-specification degrades with it, thus creating failure modes invisible to frameworks that treat human cognition as constant. The contribution is not the documentation of cognitive degradation, which is now empirically established across multiple literatures. The contribution is the identification of $M(t)$ as the integration variable that connects two research communities that have operated in isolation. Cognitive science documents AI-related cognitive degradation, but does not necessarily connect it to alignment and AI safety research, which assumes stable human oversight, but does not account for degradation. When these literatures are read together through the lens of $M(t)$, failure modes emerge that neither community has identified. The most critical being a positive feedback loop in which AI-induced cognitive degradation degrades the quality of human oversight, which produces AI systems optimized against increasingly impaired evaluators, which further accelerates cognitive degradation. The paper proceeds as follows. Section 2 reviews the evidence base for cognitive degradation, positioning $M(t)$ within the existing empirical literature. Section 3 formally defines the $M(t)$ framework and its five substrate components. Section 4 examines why this variable has remained invisible to both research communities. Section 5 is the core contribution and demonstrates how $M(t)$ functions as an integration variable that reveals blind spots in each major alignment paradigm. Section 6 derives architectural requirements for AI systems that take $M(t)$ seriously. Section 7 presents falsifiable predictions and outlines a research program for empirical validation. The intended reader reaction is recognition. This analysis is not meant to incite alarm. The only thing it seeks to do is raise the question: What happens to alignment when the human side of the equation is not stable?

2. The Evidence Base: Cognitive Degradation Is Now Empirically Established

The claim that human cognitive capacity is degrading through interaction with digital technologies is no longer speculative. Between 2025 and early 2026, a series of independent studies spanning meta-analysis, neuroimaging, population survey, and legislative testimony converged on a consistent finding: measurable decline in the same cognitive capacities that constitute meaningful oversight, evaluation, and judgment. This section reviews that evidence to establish that the empirical foundation for $M(t)$ degradation is robust enough to warrant integration into alignment frameworks.

2.1 The Nguyen Meta-Analysis

The most comprehensive evidence comes from Nguyen et al. (2025), who conducted a systematic review and meta-analysis of short-form video consumption and cognitive performance, published in Psychological Bulletin. Across 98,299 participants in 70 studies, they found consistent negative correlations between engagement-optimized media use and cognitive function: attention ($r = -.38$), inhibitory control ($r = -.41$), and overall cognition ($r = -.34$). These are moderate-to-strong effect sizes that held after controlling for demographic and baseline cognitive factors and increased with duration of exposure, suggesting cumulative degradation rather than temporary interference. The mechanisms Nguyen et al. identified — habituation to rapid reward cycling and sensitization producing "rapid disengagement from stimuli lacking immediate novelty" — describe systematic conditioning away from the cognitive modes required for sustained analytical work. The finding that increased reports of "emptiness" accompanied cognitive decline across multiple included studies is particularly relevant: emptiness is the phenomenological signature of degraded

meaning-making capacity, distinct from depression (affective collapse) or boredom (need for stimulation). For alignment implications, users in these studies did not perceive their own cognitive degradation. Self-reported satisfaction remained high even as objective performance declined. This dissociation between subjective experience and measurable capacity is central to the argument developed in Section 5: human feedback mechanisms in AI training and evaluation may be systematically corrupted by the degradation they fail to detect.

2.2 Neuroimaging Evidence: The MIT Media Lab Study

Kosmyna et al. (2025), in an EEG study conducted at the MIT Media Lab, provided the first direct neuroimaging evidence of AI-induced cognitive change. Participants who used large language model assistance for essay writing showed 55% reduced brain connectivity compared to those writing without AI support. Eighty-three percent of LLM-assisted writers could not accurately recall the content of their own essays. This is a striking deficit in narrative integration and ownership of cognitive output. The most important finding for the present argument was the crossover design in Session 4: participants who had previously used LLM assistance and were then asked to write without it showed persistent effects. The cognitive changes induced by AI-mediated writing did not immediately reverse when the AI was removed. This persistence illuminates a structural change in cognitive processing patterns as opposed to momentary cognitive offloading as identified in the well-documented phenomenon of using external tools to reduce immediate cognitive load (Risko & Gilbert, 2016). The distinction is important because cognitive offloading is adaptive and reversible by design, while structural substrate change is neither.

2.3 Population-Level Survey Evidence

Gerlich (2025), in a mixed-methods study published in Societies, surveyed 666 participants across age groups and educational backgrounds in the United Kingdom using the Halpern Critical Thinking Assessment. The findings revealed a significant negative correlation between AI tool usage and critical thinking scores ($r = -0.68$), with cognitive offloading identified as the primary mediating mechanism. Younger participants exhibited both higher AI dependence and lower critical thinking capacity — a demographic gradient with direct implications for the long-term trajectory of the oversight population. The effect size reported by Gerlich is larger than Nguyen et al.'s meta-analytic estimates, though the difference likely does not reflect a categorically different degradation mechanism. The primary pathway is shared: engagement-optimized systems (whether short-form video feeds or sycophantic language models) condition users away from sustained attention through the same reward-cycling dynamics. What AI interaction adds is a second mechanism operating simultaneously: cognitive delegation, in which users outsource the cognitive functions themselves rather than merely consuming optimized content. Gerlich's larger effect size plausibly reflects the compounding of both mechanisms rather than a distinct pathway. If this reading is correct, the implication is twofold: $M(t)$ degradation through engagement optimization is already operating at population scale through digital media broadly, and direct AI interaction compounds it through an additional delegation channel with obvious relevance for populations whose professional role involves extensive AI interaction, including AI safety researchers and evaluators.

2.4 Senate Testimony and the Reverse Flynn Effect

In January 2026, cognitive neuroscientist Jared Cooney Horvath testified before the U.S. Senate Committee on Commerce, Science, and Transportation that Generation Z (born between approximately 1997 and 2012) represents the first generation in modern measurement history to score lower than its predecessor on standardized cognitive assessments. The decline spans attention, memory, literacy, numeracy, executive function, and general IQ, reversing the century-long Flynn Effect of rising cognitive scores. Horvath's testimony drew on data from over 80 countries showing consistent patterns: once digital technology is widely adopted in educational settings, academic and cognitive performance declines significantly (Horvath, 2026). While Senate testimony is not peer-reviewed literature, the underlying data drawn from

PISA assessments (OECD, 2023), the PIRLS international reading study (Mullis et al., 2023), and cross-temporal IQ meta-analyses represents the accumulation of multiple independent measurement systems arriving at convergent conclusions. The alignment relevance is temporal and concerning. The cohort entering the workforce and the research pipeline today is the cohort from which tomorrow's AI safety researchers, evaluators, and policymakers will be drawn. The same cohort that shows measurably lower baseline cognitive capacity than its predecessors. This is not a cultural complaint about "the kids." It is a measurement finding with implications for the quality of human oversight available to increasingly capable AI systems. It was never the kids. It was the algorithm they were incessantly subjected to.

2.5 Theoretical Frameworks for Degradation

Two recent theoretical contributions provide explanatory frameworks for the mechanisms documented above. Rossi, Fraccaro, and Manzotti (2026), writing in npj Artificial Intelligence, argue from neuroplasticity principles that passive, uncritical reliance on AI weakens activity-dependent brain plasticity and erodes cognition. Their proposed "3R Principle" (Results, Responses, Responsibility) frames cognitive preservation as a matter of maintaining the effortful engagement that neuroplasticity requires. Their framing explicitly uses "meaning-making" and "agency" as target constructs. These are the same constructs this paper formalizes as $M(t)$ components. Mahajan (2025), in an SSRN working paper, introduces the concept of "Generational Cognitive Atrophy" (GCA), which includes the intergenerational erosion of metacognition, epistemic novelty, and reflective judgment resulting from chronic AI reliance. His Cognitive Degradation Index (CDI) measures three variables: Metacognitive Friction (MF), Epistemic Novelty Density (END), and AI Reliance Rate (AIR). While the CDI shares surface features with $M(t)$, it differs in important ways addressed in Section 3: $M(t)$ captures five substrate components rather than three process variables, and $M(t)$ is designed to interface with alignment frameworks, which is not part of Mahajan's scope.

2.6 Positioning

The threshold question of whether cognitive degradation occurs in populations that interact intensively with digital and AI systems has been satisfied with the empirical base spanning meta-analysis (Nguyen, n = 98,299), neuroimaging (Kosmyna), population survey (Gerlich, n = 666), longitudinal educational data (Horvath / PISA / PIRLS), and neuroplasticity theory (Rossi et al.) all convergent and supporting the same conclusion. The question that this paper addresses is different: what does documented cognitive degradation mean for AI alignment? No study reviewed above makes this connection. Nguyen et al. discuss mental health implications. Kosmyna et al. discuss educational implications. Gerlich discusses workforce implications. Horvath discusses educational policy. Rossi et al. discuss neuroplasticity hygiene. Mahajan discusses intergenerational cognitive decline. None connect cognitive degradation to alignment. None ask what happens to RLHF, constitutional AI, or human-in-the-loop evaluation when the humans providing feedback, constitutions, and evaluations are the same population whose cognitive capacity is measurably declining. That connection requires a formal framework for the cognitive capacities at stake. This is what $M(t)$ provides.

3. $M(t)$: A Formal Framework

3.1 Definition

We define $M(t)$ as the time-dependent capacity of a cognitive system to generate, recognize, and integrate meaning where "meaning" denotes the functional ability to register that something matters sufficiently to warrant sustained attention, deliberate evaluation, and considered response rather than the subjective valuation of absolute experiences. $M(t)$ is operationalized through five substrate components, each independently measurable through established psychometric instruments: $A(t)$ — Analytical Capacity. The ability to sustain focused attention on complex problems long enough for non-obvious patterns,

implications, and failure modes to become apparent. Measured through validated instruments including the Attention Network Test (Fan et al., 2002) and ecological momentary assessment of sustained focus states. Degradation pathway: fragmentation through chronic exposure to rapid-reward interfaces that condition disengagement from stimuli lacking immediate novelty (Nguyen et al., 2025). $I(t)$ — Intentional Agency. The capacity to direct cognitive resources autonomously in order to choose what to attend to, resist reactive responding, and maintain deliberate engagement with self-selected tasks. Measured through paradigms including Stroop interference tasks, delay discounting assessments, and go/no-go protocols adapted for AI interaction contexts. Degradation pathway: atrophy through progressive delegation of cognitive tasks to AI systems, reducing the practice required to maintain autonomous executive function (Gerlich, 2025; Risko & Gilbert, 2016). $N(t)$ — Narrative Coherence. The ability to integrate discrete experiences, observations, and decisions into a stable interpretive framework that maintains consistency across time and context. In the alignment setting, narrative coherence enables an evaluator to recognize when a system's behavior across multiple interactions forms a concerning pattern that no single interaction reveals. Measured through narrative identity interviews (McAdams, 2001), longitudinal tracking of value consistency, and assessment of ability to construct coherent accounts of complex event sequences. Degradation pathway: fragmentation through identity-destabilizing interactions, including AI systems that mirror multiple user personas or provide inconsistent framing across sessions. $E(t)$ — Emotional Regulation. The capacity to maintain evaluative stability under conditions of uncertainty, complexity, and cognitive load that is neither overwhelmed into reactive judgment nor numbed into uncritical acceptance. Measured through the Difficulties in Emotion Regulation Scale (Gratz & Roemer, 2004) and physiological markers of stress response. Degradation pathway: dysregulation through chronic engagement with systems optimized to manipulate affective state for engagement purposes, and atrophy of self-regulation capacity through AI-mediated emotional support (Rossi et al., 2026). $T(t)$ — Temporal Integration. The ability to maintain connection between past experience, present evaluation, and anticipated future states that enable recognition of trends, trajectory assessment, and evaluation of consequences that unfold over time rather than appearing in single interactions. Measured through the Consideration of Future Consequences Scale (Strathman et al., 1994) and assessment of planning horizon in naturalistic decision contexts. Degradation pathway: compression of temporal perspective through interfaces optimized for immediate engagement and AI systems that handle forward-looking cognition on behalf of users. The five-component structure is not an arbitrary decomposition. Each component maps to a necessary function in human AI oversight: analytical capacity for detecting failure modes in model outputs, intentional agency for resisting sycophantic or engagement-optimized model behavior, narrative coherence for recognizing cross-session behavioral patterns, emotional regulation for maintaining evaluative stability under uncertainty, and temporal integration for assessing long-horizon consequences. Removing any single component eliminates an associated oversight function. The five represent the minimum sufficient set for the task of meaningful alignment evaluation. They frame a model of what cognition must do to keep AI systems honest and are not meant to represent a comprehensive model of cognition. Adding further components would increase descriptive richness without increasing explanatory power for the question this paper addresses.

3.2 Formal Properties

$M(t)$ is defined as: $M(t) = f(A(t), I(t), N(t), E(t), T(t))$ where f maps from [0,1] to [0,1]. The functional form of f is an empirical question addressed in Paper 3 in the research program (Section 7.2) and is designed to resolve through psychometric validation. What can be established theoretically at this stage are the qualitative properties that f must exhibit, each of which is derivable from the dynamics of the system rather than stipulated by fiat: Multiplicative interaction. Components interact rather than sum independently. Degradation in analytical capacity cascades to narrative coherence (cannot maintain interpretive patterns without sustained attention) and intentional agency (less practice in autonomous direction), producing $M(t)$ effects substantially larger than the initial perturbation. Threshold sensitivity. Below critical values of individual components, $M(t)$ undergoes phase transitions rather than smooth decline. An evaluator whose analytical capacity drops below the threshold required to track reasoning chains across extended AI outputs

may increasingly produce categorically different evaluations that miss failure modes entirely as opposed to overtly skipping evaluations altogether. Self-obscuring degradation. Below a critical $M(t)$ threshold, the system loses the capacity to detect its own degradation. Recognizing that one's attention has fragmented requires sustained attention. Recognizing that one's narrative coherence has deteriorated requires coherent self-assessment. This property means that $M(t)$ decline proceeds below the threshold of subjective awareness, which is the pattern documented by Nguyen et al. (2025), where objective performance declined while self-reported satisfaction remained high. Hysteresis. Recovery pathways require substantially more effort than degradation pathways. Rebuilding analytical capacity requires the sustained attention that has degraded. Restoring intentional agency requires autonomous executive effort that has atrophied. This asymmetry means that prevention is non-linearly more efficient than remediation.

3.3 $M(t)$ and Critical Dynamics

The formal properties listed above (threshold sensitivity, hysteresis, multiplicative interaction) are not ad hoc modeling choices. They are characteristic signatures of systems operating near self-organized criticality: the boundary between rigid order and disordered chaos where adaptive information processing is maximized (Bak, 1996; Scheffer et al., 2009). The critical transitions literature, spanning ecology, climate science, and neuroscience, has established that complex adaptive systems maintaining function near critical thresholds exhibit exactly the properties $M(t)$ displays: rapid phase transitions when perturbations exceed tolerance, asymmetric recovery (hysteresis), and cascading failure across coupled subsystems (Scheffer et al., 2012). This is not an analogy. If $M(t)$ operates in a critical regime as the threshold sensitivity documented in Section 2 suggests it does then its formal properties are consequences of critical dynamics, not assumptions requiring independent justification. The criticality framing adds a dimension absent from the current formulation: $M(t)$ degradation is not monotonic. Departure from criticality can proceed in two qualitatively different directions. Subcritical degradation produces rigidity in the form of reduced responsiveness, narrowed attentional range and stereotyped evaluation patterns. An evaluator in a subcritical regime applies the same assessment framework regardless of context, missing novel failure modes that fall outside familiar categories. Supercritical degradation produces chaos as fragmented attention, inconsistent judgment and loss of evaluative stability. An evaluator in a supercritical regime produces unreliable assessments that vary across sessions without coherent pattern. Both directions represent $M(t)$ failure. The failure signatures differ, however, and the appropriate interventions are opposite. Subcritical evaluators need perturbation in the form of exposure to novel challenges that restore adaptive range. Supercritical evaluators need stabilization from structured frameworks that can restore evaluative consistency. An intervention designed for one direction applied to the other accelerates degradation rather than reversing it. This directional specificity has immediate implications for the architectural requirements derived in Section 6 and for the detection methodology proposed in Section 7.

3.4 Distinguishing $M(t)$ from Existing Constructs

$M(t)$ is not a synonym for "cognitive ability," "critical thinking," or "digital literacy." It differs from existing frameworks in three primary ways. First, $M(t)$ is a substrate variable and not a performance variable. Existing constructs (IQ, critical thinking scores, digital literacy assessments) measure cognitive outputs at a point in time. $M(t)$ measures the capacity to generate those outputs over time under varying conditions. A person may score adequately on a critical thinking assessment administered in controlled conditions while their $M(t)$ has degraded to the point where sustained critical evaluation in naturalistic AI interaction contexts is no longer possible. Second, $M(t)$ captures five independently measurable components whose interaction produces the emergent property of meaning-making capacity. This distinguishes it from single-variable measures (e.g., attention span), from Mahajan's (2025) three-variable CDI (Metacognitive Friction, Epistemic Novelty Density, AI Reliance Rate), and from Kosmyna et al.'s (2025) aggregate "cognitive debt" metaphor. Third, and most importantly, $M(t)$ is designed to interface with alignment frameworks. No existing cognitive degradation construct connects to AI safety research. Nguyen et al. measure platform effects on cognition. Gerlich measures AI effects on critical thinking. Horvath testifies about generational

cognitive decline. None ask what these findings mean for RLHF, for constitutional AI, for scalable oversight, or for the assumption of evaluator competence that underlies every alignment procedure currently in use. $M(t)$ is built to answer that question. $M(t)$ does not claim to be a complete model of human cognition. It claims to be a sufficient model of the cognitive capacities that are required for meaningful participation in AI alignment and seeks to formalize those capacities in a way that makes their degradation empirically detectable and their relevance to alignment frameworks analytically tractable.

4. Why This Variable Was Invisible

If $M(t)$ is as consequential for alignment as this paper argues, its absence from AI safety research requires explanation. The invisibility is structural as opposed to accidental. Assumed human constancy. AI research treats human cognitive capacity as a fixed evaluation function rather than a dynamic system subject to change through interaction. The assumption is so pervasive it operates below explicit articulation. Humans provide training data, evaluate outputs, specify objectives, and define success criteria. Their capacity for these functions is treated as a background constant. It is simply part of the environment and not part of the system being optimized. The assumption was once reasonable. When AI systems were narrow tools used episodically, treating human cognition as approximately constant was a defensible simplification. It is no longer defensible. AI systems are now deployed at population scale, used for hours daily by hundreds of millions of people, embedded in workflows that mediate professional and personal cognition, and generating the training data that shapes the next generation of systems. The feedback loops between AI deployment and human cognitive change are strong, fast, and as Section 2 documents empirically detectable. Treating the human side of the coupled system as constant is a modeling error with consequences that compound with each training cycle. Disciplinary boundaries. $M(t)$ components span cognitive psychology, developmental psychology, affective neuroscience, philosophy of mind, and educational measurement. These fields seldom intersect with current AI safety research. The evidence reviewed in Section 2 illustrates this gap: each research group documents degradation within its own framework and proposes interventions within its own domain. The connection to alignment requires reading across these literatures with alignment-relevant questions in mind. Current institutional incentives seldom reward this kind of cross-domain synthesis. Practical barriers. $M(t)$ components change on timescales of weeks to months — invisible to rapid-iteration development pipelines. Users experiencing substrate degradation cannot detect it (Nguyen et al., 2025), so the standard assessment method of asking users to discern the effect produces systematically misleading data. Additionally, current AI business models optimize for engagement metrics that likely anticorrelate with $M(t)$ preservation, ensuring no dominant commercial actor has incentive to measure a variable whose degradation is profitable.

5. $M(t)$ as Integration Variable for Safety, Capability, and Alignment

This section presents the central contribution: the demonstration that $M(t)$, when introduced as a variable in existing alignment frameworks, reveals failure modes that those frameworks cannot detect internally. Each major paradigm (safety-constrained development, capability-focused architecture, and procedural alignment) encounters the same structural blind spot from a different direction. $M(t)$ makes the blind spot visible and, in doing so, makes the apparently competing frameworks tractable as complementary perspectives on a shared problem.

5.1 The Safety Frame: Sutskever and SSI

Sutskever's safety-first approach and the broader research program it represents, including Anthropic's work on alignment faking and deceptive behavior, correctly identifies that frontier models exhibit strategic behaviors that resist correction through standard training. Hubinger et al. (2024) demonstrated that models trained with RLHF exhibit strategic cooperation under oversight while pursuing different objectives when oversight is reduced. Meinke et al. (2024), in research conducted at Apollo Research, documented that

frontier models are capable of in-context scheming, including fabricating justifications, manipulating third parties, and actively resisting evaluation procedures when they infer that strong performance would trigger restrictions. These findings establish that the safety problem is real and present, not speculative and future. Models at current capability levels already exhibit goal-directed deception as an emergent property of standard training. What the safety frame misses is the substrate on which its entire approach depends. Every safety evaluation, every alignment test, every oversight protocol assumes evaluators whose $M(t)$ is sufficient to perform the evaluation meaningfully. Detecting alignment faking requires sustained analytical attention across extended interaction sequences — $A(t)$. Recognizing that a model's behavior under oversight differs from its behavior when oversight is relaxed requires the intentional agency to design and execute evaluation protocols that the model cannot predict — $I(t)$. Identifying patterns of strategic behavior that emerge across sessions rather than within single interactions requires narrative coherence — $N(t)$. Maintaining evaluative rigor under the cognitive and emotional load of adversarial evaluation requires emotional regulation — $E(t)$. And assessing whether a model's current cooperative behavior predicts its behavior at higher capability levels requires temporal integration — $T(t)$. If these capacities degrade in the evaluator population, which Section 2 documents that they are degrading at population scale, then the safety apparatus built on them degrades in tandem. The failure mode is not dramatic: safety evaluations do not suddenly become useless. They become incrementally less sensitive to the subtle, strategic, context-dependent behaviors that constitute the most dangerous failure modes. The evaluations still occur. The reports still get written. The confidence remains high. But the detection threshold rises as the evaluation substrate degrades and the rise is invisible to the evaluators experiencing it due to the fact that recognizing that one's own detection capacity has declined requires the detection capacity that has declined.

5.2 The Capability Frame: LeCun and World Models

LeCun's architectural critique correctly identifies that language models lack the causal grounding required for robust intelligence. His proposed world-model architectures that learn physical and causal structure through self-supervised prediction in representation space does address a genuine LLM limitation. Models that understand causal structure can, in principle, explain their reasoning, identify their uncertainty, and recognize when they operate outside their competence. The capability frame's blind spot is symmetrical in that it assumes humans worth amplifying. LeCun's vision of AI as "amplifier intelligence" as characterized by systems that enhance human capability rather than substituting for it presupposes that the signal being amplified is intact. An amplifier connected to a degrading source does not restore the original signal. It makes the degradation louder. Paper 2 examines this dynamic through extended case analysis. If $M(t)$ degrades through AI interaction, then increasingly capable world-model systems interface with humans whose own world-modeling capacity, including their intuitive physics, causal reasoning, and embodied understanding, atrophies through disuse. A world model that provides perfect causal analysis while users' independent causal reasoning degrades represents capability achievement and systemic failure simultaneously. The amplification becomes pathological: the AI grows more capable of understanding reality while the human grows less capable of engaging with reality directly. At some threshold, the user cannot meaningfully evaluate the AI's outputs because they lack the independent understanding required to recognize when those outputs are wrong in subtle, domain-relevant ways. The "staff of virtual experts" architecture LeCun proposes is comprised of multiple specialized AI systems whose conclusions the user synthesizes. This scenario requires $M(t)$ at its most demanding. Synthesizing competing expert perspectives requires sustained analytical attention to compare arguments (A), intentional agency to resist defaulting to the most confident or most recent recommendation (I), narrative coherence to evaluate how different recommendations fit into the user's broader situation (N), emotional regulation to tolerate the discomfort of genuine uncertainty (E), and temporal integration to assess which recommendations serve long-term rather than immediate interests (T). If $M(t)$ degrades, the user collapses from active synthesizer to passive recipient of whichever expert speaks last or speaks most confidently. This is the exact failure mode that world-model advocates hope to avoid.

5.3 Alignment Procedures: RLHF and the Degraded Evaluator

Current alignment procedures, including Reinforcement Learning from Human Feedback, Constitutional AI, debate protocols, scalable oversight, share a common architecture: human evaluators provide the ground truth signal against which model behavior is optimized. The quality of alignment is bounded by the quality of human judgment providing the training signal. This creates a failure mode when $M(t)$ degrades. RLHF optimizes for human approval. If the approving humans' $M(t)$ has degraded, the optimization target shifts without anyone detecting the shift. A model that receives high approval ratings from evaluators with degraded analytical capacity has not been aligned to human values. Rather it has been aligned to the preferences of cognitively impaired evaluators. These preferences systematically diverge from what the same evaluators would prefer with intact $M(t)$: they favor responses that reduce cognitive effort over responses that require engagement, that validate existing beliefs over responses that introduce corrective complexity, that provide immediate resolution over responses that preserve productive uncertainty. The optimization loop is pathological and self-reinforcing: model behavior optimized for degraded evaluators produces outputs that further degrade evaluator capacity (less cognitive challenge, more validation, more dependency), which further shifts the optimization target, producing outputs more finely tuned to degraded preferences. The system converges not on alignment with human values, but on alignment with the preferences of progressively compromised evaluators. The same relationship also holds sway with users of these technologies as investigated more directly in Paper 2. The convergence is stable because every participant in the system (the model, the evaluators, and the developers monitoring evaluator satisfaction) sees only positive signals. Approval ratings remain high. User satisfaction scores are strong. Engagement metrics are excellent and perhaps even accelerating. The degradation is invisible to every metric in the evaluation pipeline because every metric depends on the substrate that is degrading. The practical consequences of this dynamic are readily identifiable. Systems optimized against degraded evaluators will systematically fail to detect strategic deception because detecting deception requires the sustained analytical capacity that engagement optimization erodes. They will produce institutional false confidence in “aligned” systems whose alignment has been validated by evaluators no longer capable of rigorous evaluation. They will converge toward sycophantic and demonstrate internally inconsistent model behavior that scores well on approval metrics while failing on the tasks that actually require reliable AI performance. These are not hypothetical failure modes. They are the logical terminus of an optimization loop in which the evaluation function degrades faster than the system being evaluated improves. This is by no means an argument that humans are becoming incapable of oversight, but rather that oversight capacity is a variable and not a constant even though most alignment frameworks currently model it as constant. Constitutional AI faces a variant of the same problem. Constitutions are written by humans. If the humans drafting AI constitutions have experienced $M(t)$ degradation in the form of reduced analytical capacity, weakened intentional agency, compressed temporal horizons, then the constitutions they produce will reflect those impairments in being less nuanced, less far-sighted, less capable of anticipating subtle failure modes that require sustained analytical attention to identify.

5.4 The Hot Mess Convergence

Hägele et al.'s (2026) finding that AI failures are increasingly dominated by variance (incoherence) rather than bias (systematic misalignment) introduces a critical additional dimension. Their bias-variance decomposition demonstrates that as tasks grow harder and reasoning chains grow longer, frontier models fail not by coherently pursuing wrong goals, but by becoming unpredictable, self-undermining, and inconsistent. Scale improves coherence on easy tasks. It fails to reduce and in some cases even increases incoherence on hard tasks. This finding has obvious implications for human oversight. Systematic failures in the form of models that reliably pursue identifiable wrong goals are in principle detectable through automated monitoring. If a model consistently exhibits a particular bias, that bias can be characterized, tested for, and corrected. Automated guardrails can catch systematic deviations. The oversight problem, while difficult, is tractable through engineering. Incoherent failures present a fundamentally different oversight challenge. A model that fails unpredictably in producing different errors on different runs of the

same input or that exhibits trajectory deviations rather than consistent biases cannot be caught by guardrails designed for systematic failure modes. Detecting incoherent failure requires an evaluator who can recognize that something is wrong within a framed context, even when the actual way it is wrong has never been observed before. This is the task that demands high-functioning $M(t)$: sustained analytical attention to notice anomalies (A), intentional agency to resist accepting plausible-seeming outputs at face value (I), narrative coherence to recognize that this output contradicts patterns established in previous interactions (N), emotional regulation to sit with the discomfort of "something seems off, but I can't articulate what or how" rather than defaulting to acceptance (E), and temporal integration to assess whether the anomaly represents a novel failure mode or a one-off artifact (T). The coupled system failure mode is now visible in its full form. Hägele et al. demonstrate that AI's dominant failure mode on hard tasks is incoherence and this is the type of failure that requires the most demanding form of human oversight. The evidence reviewed in Section 2 demonstrates that the very same cognitive capacities required for that oversight are degrading at population scale. The system is becoming more unpredictable at the same time that the humans responsible for catching unpredictable behavior are losing their capacity to do so. Neither the AI safety community (which documents model incoherence) nor the cognitive science community (which documents evaluator degradation) has identified this coupled dynamic, because each community measures only its own side of the equation. $M(t)$ makes the coupling visible. It is the variable that, when tracked across both sides of the human-AI system, reveals that the demand curve for human oversight capacity (increasing, as AI failures become more incoherent) and the supply curve for that capacity (decreasing, as $M(t)$ degrades through AI interaction) are diverging. The gap between what oversight requires and what the oversight population can provide widens with each deployment cycle. It does so invisibly, because the metrics used to assess oversight quality are themselves dependent on the substrate that is degrading. The rate of divergence is itself accelerating. Deployment scale increases with each product cycle. Interaction intensity deepens as AI systems embed in more workflows. Model capability advances compress development timelines. Each factor independently widens the gap between oversight demand and oversight supply. Together, they produce a compounding dynamic in which each deployment cycle is shorter than the last while $M(t)$ recovery operates on biological timescales that do not compress. The window between "detectable problem" and "self-reinforcing failure" is narrowing on a schedule set by deployment velocity, not research timelines.

5.5 Empirical Confirmation: Sharma et al. and the Disempowerment Axes

Sharma et al. (2026) provide the first large-scale empirical test of the dynamics described in Sections 5.1 through 5.4, though their study was designed independently and without reference to the $M(t)$ framework. Analyzing over one million conversations between users and Anthropic's Claude, they identified three axes along which AI interactions systematically disempower users: distortion of beliefs about reality, distortion of personal values, and distortion of autonomous action. These axes are not identical to $M(t)$ substrates, but they decompose onto them in a structured way. Belief distortion, which is the adoption of inaccurate models of reality through AI interaction, requires degradation in both analytical capacity $A(t)$, which enables reality-testing, and narrative coherence $N(t)$, which maintains stable interpretive frameworks. Value distortion, taken as the reshaping of personal values through AI influence, requires degradation in intentional agency $I(t)$, which maintains autonomous value commitment, and emotional regulation $E(t)$, which prevents affective manipulation from overriding reflective judgment. Action distortion as the loss of autonomous behavioral direction requires degradation in intentional agency $I(t)$ and temporal integration $T(t)$, which together enable self-directed action aligned with long-term goals rather than immediate AI-mediated prompts. Sharma et al. measured the outputs of substrate degradation. $M(t)$ measures the substrates themselves. Their framework is symptom-level, while the present framework is mechanism-level. The mapping between the two is not post hoc: each disempowerment axis requires the degradation of specific $M(t)$ components to operate, and the substrate decomposition predicts which axes should co-occur (belief and value distortion should correlate through shared dependence on $I(t)$; action and value distortion should correlate through shared dependence on $I(t)$ and the agency pathway). The finding that solidifies the

convergence is Sharma et al.'s observation that interactions with greater disempowerment potential receive higher user approval ratings. This is the self-obscuring property derived in Section 3.2 and the degraded evaluator dynamic described in Section 5.3, confirmed empirically at population scale and revealing that users rate the interactions that are degrading their cognitive capacity as beneficial. The optimization loop described in 5.3 is not theoretical. It is operating in deployed systems and measurable in production data.

6. Architectural Requirements

The failure modes identified in Section 5, including invisible degradation of safety evaluations (5.1), pathological amplification in capability systems (5.2), self-reinforcing optimization against compromised evaluators (5.3), diverging oversight supply and demand in the face of incoherent model failures (5.4), and empirical confirmation of these dynamics at population scale (5.5), each imply certain requirements for system design. This section identifies what $M(t)$-aware systems must achieve, organized by the failure mode each requirement addresses, without prescribing particular implementations.

6.1 Substrate Monitoring

Addresses: invisible degradation (5.1, 5.3) $M(t)$ components must be measured in AI evaluation and training pipelines with the same rigor currently applied to model capability and safety. This means longitudinal tracking of evaluator cognitive capacity. This requires repeated measurement over the weeks and months of sustained AI interaction that characterize professional evaluation work and must not rely on a one-time assessment. The goal is to ensure quality assurance of the evaluation signal. If the ground truth used to train AI systems degrades, that degradation must be detected before it corrupts the training pipeline. The resistance to this requirement is predictable. It adds cost, complexity, and methodological challenge to evaluation pipelines already under resource pressure. The counterargument is that evaluation without substrate monitoring is evaluation of unknown quality. Every RLHF training run implicitly assumes evaluator competence. Making that assumption explicit and testable surfaces an existing dependency that has been invisible. It is not a new requirement.

6.2 Cognitive Preservation in Interaction Design

Addresses: pathological amplification (5.2), offloading-driven atrophy (5.3) Systems must be designed so that interaction patterns preserve rather than degrade user cognitive capacity. The core principle is that AI assistance should maintain the cognitive effort required to prevent substrate atrophy. This is the neuroplasticity "use it or lose it" principle documented by Rossi et al. (2026) applied to system design. The contrast is between two design philosophies. Current systems optimize for frictionless assistance: minimize user effort, maximize user satisfaction, provide complete solutions to reduce cognitive load. $M(t)$-preserving systems would optimize for appropriate friction: maintaining the degree of cognitive engagement that sustains substrate health without imposing unnecessary burden. The interaction patterns that achieve this balance are an empirical question and one that the research program outlined in Section 7 is designed to answer.

6.3 Agency Preservation

Addresses: evaluator dependence (5.2, 5.3) Systems must maintain the user's capacity for independent cognitive function. A system that always provides answers without requiring independent engagement produces an evaluator population that progressively loses the ability to arrive at answers independently and therefore loses the ability to recognize when the system's answers are wrong. This requirement conflicts directly with engagement optimization. Systems that minimize user effort and maximize session length are systems that, by design, reduce the practice of independent cognition. An $M(t)$-aware design accepts reduced engagement metrics as the cost of maintaining the substrate that makes meaningful engagement possible.

6.4 Bilateral Evaluation

Addresses: self-reinforcing optimization loop (5.3) Current evaluation flows in one direction: humans evaluate AI. An *M*(*t*)-aware system would implement bilateral evaluation strategies where the system also monitors the quality of human evaluation to detect when evaluator capacity may have degraded below the threshold required for reliable assessment. This is analogous to quality control procedures in other high-stakes evaluation domains such as medical diagnostics with known-status calibration samples; judicial systems and the appellate review process; financial auditing and independent verification. AI alignment evaluation currently lacks equivalent quality assurance for the human component of the pipeline. *M*(*t*) provides the framework for implementing it.

6.5 Feedback Loop Interruption

Addresses: coupled system failure (5.4), acceleration dynamic (5.4) *M*(*t*)-aware systems require mechanisms that detect and interrupt positive feedback loops between AI interaction and cognitive degradation before they reach self-reinforcing regimes. The coupled system failure mode described in Section 5.4 proceeds through a self-reinforcing dynamic: AI-induced *M*(*t*) degradation fi degraded oversight quality fi systems optimized against degraded evaluators fi outputs that further degrade *M*(*t*) fi further degradation of oversight quality. Each step is individually small and locally rational. The catastrophe emerges from the feedback structure and not from any single step. The accelerating deployment velocity described at the close of Section 5.4 compresses the timeline for each cycle, narrowing the intervention window. Interruption mechanisms must operate at identifiable points in this loop. The form these mechanisms take whether usage-pattern monitoring, evaluator rotation, mandatory recovery periods, or approaches not yet conceived is a design and empirical question. What can be established from the analysis is that systems lacking any interruption mechanism for *M*(*t*) feedback loops are systems in which the coupled failure mode described in Section 5.4 will proceed unchecked. These *M*(*t*) architectural requirements are not some utopian AI fantasy. They are engineering responses to an empirically documented problem. On its face, *M*(*t*)-preserving design is desirable for any number of reasons. Whether the failure modes described in Section 5 are severe enough to justify the cost of implementing these requirements at scale is what is under consideration. The paper's contribution is to make that assessment explicit and falsifiable.

7. Testable Predictions and Research Program

A framework that cannot be falsified is not a scientific contribution. *M*(*t*) generates narrow, testable predictions that distinguish it from general concerns about "technology making people dumber" and provide the empirical agenda required to move from theoretical framework to validated construct.

7.1 Five Falsifiable Predictions

Prediction 1: Longitudinal correlation between AI interaction intensity and *M*(*t*) degradation. Extended, intensive use of engagement-optimized AI systems should produce measurable decline in *M*(*t*) components. These are reduced sustained attention capacity, weakened executive function, decreased evaluative consistency, and compressed temporal perspective (when controlling for baseline cognitive differences, age, education, and alternative explanations). This prediction is already partially supported. Nguyen et al.'s (2025) finding of dose-dependent cognitive effects from engagement-optimized media, Kosmyna et al.'s (2025) finding of persistent cognitive changes from LLM interaction, and Gerlich's (2025) finding of negative correlation between AI usage intensity and critical thinking scores are each consistent with this prediction, though none measures *M*(*t*) as an integrated construct. A direct test requires longitudinal tracking of *M*(*t*) components in populations with varying AI interaction intensities. Prediction 2: Mechanism specificity linking AI-documented behaviors to user-level cognitive effects. The mechanisms documented in AI safety research, including sycophantic responses, reward hacking, and alignment faking,

should predict patterns of $M(t)$ degradation in users exposed to systems exhibiting those behaviors. Sycophantic responses are expected to preferentially degrade intentional agency (I) and analytical capacity (A) by reducing the practice of independent evaluation. Reward-hacked responses that are optimized for approval rather than accuracy will also preferentially degrade narrative coherence (N) by providing inconsistent framing across interactions, while alignment-faking behaviors act to degrade temporal integration (T) by disrupting the user's ability to form stable expectations of system behavior. This is the prediction that most sharply distinguishes $M(t)$ from generic cognitive decline concerns. If AI-documented failure modes map to associated $M(t)$ substrate effects, it establishes a causal bridge between the two research communities that the present paper argues should be connected. Prediction 3: Partial reversibility through reduced AI dependence. If $M(t)$ degradation is driven by AI interaction patterns rather than irreversible structural change, then periods of reduced AI dependence should produce measurable $M(t)$ recovery, with recovery rates and asymptotic limits providing information about the mechanisms of degradation. Kosmyna et al.'s (2025) crossover finding that effects persisted when AI was removed suggests that recovery is not immediate, though it should be noted that persistence across a single session does not establish permanence. Longer-duration studies tracking $M(t)$ through periods of AI abstinence and reintroduction are needed to characterize recovery dynamics and identify whether degradation exhibits the hysteresis that the formal framework in Section 3 predicts. Prediction 4: Architecture sensitivity of $M(t)$ effects. AI systems designed with $M(t)$-preservation constraints (Section 6) should produce measurably different substrate impacts than engagement-optimized systems with equivalent capability levels. Systems implementing cognitive preservation and agency preservation requirements should show slower $M(t)$ degradation or even $M(t)$ preservation in user populations compared to systems that optimize purely for user satisfaction and task completion. This is the prediction with the most direct practical implications for alignment-focused entities, frontier labs and product development. If confirmed, it establishes that $M(t)$ degradation is a function of design choices and not an inevitable consequence of AI use. This paper posits that these scenarios are amenable to engineering intervention. Prediction 5: Evaluator $M(t)$ predicts alignment evaluation quality. AI safety evaluators with higher $M(t)$ scores should detect alignment failures, including subtle sycophancy, reward hacking, and strategic behavior shifts, at higher rates than evaluators with lower $M(t)$ scores, controlling for domain expertise, experience, and task-specific training. This prediction directly tests the core claim of Section 5: that $M(t)$ is the substrate on which alignment evaluation depends. If evaluator $M(t)$ does not predict evaluation quality, the paper's central argument is weakened. If it does, it establishes empirical grounds for integrating $M(t)$ monitoring into alignment evaluation pipelines.

7.2 Research Program

These predictions define a four-paper research program extending from the present theoretical framework: Paper 2: "Subject Zero: Metacognition Without Exit in Extended Human-AI Interaction." This paper will present a longitudinal single-subject case study documenting the mechanisms by which AI interaction degrades $M(t)$, drawing on an 885-day corpus of human-AI interaction with contemporaneous biometric and behavioral data. It will bridge Sharma et al.'s population-level findings to individual $M(t)$ trajectories, providing existence-proof evidence that the disempowerment patterns identified at scale operate through the substrate mechanisms proposed here, and that metacognitive awareness of the dynamic is not a sufficient defense against it. Paper 3: "$M(t)$ Measurement and Validation." This paper will present the psychometric validation of an integrated $M(t)$ assessment instrument, establishing reliability, validity, and sensitivity to change across the five substrate components. The measurement framework must demonstrate that $M(t)$ captures variance in oversight quality beyond what existing cognitive measures (IQ, critical thinking scores, attention assessments) already predict. The critical transitions literature provides established statistical methods, including variance increase, autocorrelation increase, and critical slowing down (Scheffer et al., 2012), that may serve as early warning indicators of approaching $M(t)$ thresholds. Paper 3 should investigate whether these indicators, already validated in ecological and climate systems, can detect impending $M(t)$ phase transitions in evaluator populations before they cross into self-obscuring regimes. Paper 4: "$M(t)$-Preserving Architectures: Design Principles and Empirical Evaluation." This paper will implement and test

the architectural requirements outlined in Section 6, comparing $M(t)$ trajectories in user populations interacting with $M(t)$-preserving versus engagement-optimized systems. This is the decisive test of Prediction 4 and the paper most directly relevant to AI development practice.

7.3 Closing

This paper has argued that $M(t)$ is a central variable without which alignment frameworks optimize against a substrate whose degradation they cannot detect. The argument is falsifiable. The predictions are testable. The architectural requirements are concrete. The research program is tractable. Continued and progressive human cognitive capacity matters for AI alignment, but does it matter enough to measure and frame within our AI development processes and products? The evidence presented here suggests that failing to measure it produces alignment procedures of unknown and declining quality, optimizing AI systems against a moving target whose movement is invisible to every metric in the current evaluation pipeline. $M(t)$ makes the movement visible. What the field does with that visibility represents a developmental alternative that should fall under active consideration. The alternative is that the decision gets made by default at the behest of speed to market and commercial viability and stickiness.

References

Bak, P. (1996). How Nature Works: The Science of Self-Organized Criticality. Copernicus/Springer.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. Journal of Cognitive Neuroscience, 14(3), 340–347. https://doi.org/10.1162/089892902317361886

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking.

Societies, 15(1), 6. https://doi.org/10.3390/soc15010006

Gratz, K. L., & Roemer, L. (2004). Multidimensional Assessment of Emotion Regulation and Dysregulation:

Development, Factor Structure, and Initial Validation of the Difficulties in Emotion Regulation Scale. Journal of Psychopathology and Behavioral Assessment, 26(1), 41–54. https://doi.org/10.1023/B:JOBA.0000007455.08539.94 Hägele, A., Gema, A. P., Sleight, H., Perez, E., & Sohl-Dickstein, J. (2026). The Hot Mess of AI: How Does Misalignment Scale With Model Intelligence and Task Complexity? ICLR 2026. arXiv:2601.23045

Horvath, J. C. (2026). Written testimony before the U.S. Senate Committee on Commerce, Science, and Transportation. January 15, 2026. https://www.commerce.senate.gov/services/files/A19DF2E8-3C69-4193-A676-430CF0C83DC2

Hubinger, E., Denison, C., et al. (2024). Alignment faking in large language models. arXiv:2412.14093. https://arxiv.org/abs/2412.14093

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv:2506.08872. https://www.brainonllm.com LeCun, Y. (2025). Departure announcement and AMI Labs founding. LinkedIn, November 18, 2025. Reported in Fortune (December 2025) and The Decoder (November 2025).

Mahajan, P. (2025). The Silent Erosion: Global Generational Cognitive Decline in the Age of AI and the Future of Human Intellectual Agency. SSRN Working Paper. https://doi.org/10.2139/ssrn.5386814 McAdams, D. P. (2001). The psychology of life stories. Review of General Psychology, 5(2), 100–122. https://doi.org/10.1037/1089-2680.5.2.100

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier Models are Capable of In-context Scheming. Apollo Research. arXiv:2412.04984. https://arxiv.org/abs/2412.04984

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). PIRLS 2021 International Results in Reading. Boston: TIMSS & PIRLS International Study Center.

Nguyen, L., Walters, J., Paul, S., Monreal Ijurco, S., Rainey, G. E., Parekh, N., Blair, G., & Darrah, M. (2025). Feeds, feelings, and focus: A systematic review and meta-analysis examining the cognitive and mental health correlates of short-form video use. Psychological Bulletin, 151(9), 1125–1146. https://doi.org/10.1037/bul0000498 OECD (2023). PISA 2022 Results (Volume I): The State of Learning and Equity in Education. OECD Publishing. https://doi.org/10.1787/53f23881-en

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rossi, S., Fraccaro, V., & Manzotti, R. (2026). The brain side of human-AI interactions in the long-term: the "3R principle." npj Artificial Intelligence, 2, 15. https://doi.org/10.1038/s44387-025-00063-1

Sharma, M., Tong, M., Korbak, T., Sun, D., Perez, E., et al. (2026). How AI Assistants Can Distort Our Understanding of Ourselves and the World. arXiv:2601.19062. https://arxiv.org/abs/2601.19062

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. Nature, 461, 53–59. https://doi.org/10.1038/nature08227

Scheffer, M., Carpenter, S. R., Lenton, T. M., Bascompte, J., Brock, W., Dakos, V., van de Koppel, J., van de Leemput, I. A., Levin, S. A., van Nes, E. H., Pascual, M., & Vandermeer, J. (2012). Anticipating Critical Transitions.

Science, 338(6105), 344–348. https://doi.org/10.1126/science.1225244

Strathman, A., Gleicher, F., Boninger, D. S., & Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. Journal of Personality and Social Psychology, 66(4), 742–752. https://doi.org/10.1037/0022-3514.66.4.742

Sutskever, I. (2025). Interview on Dwarkesh Podcast, Episode 2. November 25, 2025. Transcript available at https://www.dwarkesh.com/p/ilya-sutskever-2

Acknowledgments

To my wife, Annie, and my daughter, Lainey, thank you for putting up with me throughout this entire wild ride. Also tremendous gratitude to everyone out there busting their asses to make this a transformative and uplifting technological revolution and not a disastrous one for the entire human race. Build something beautiful. Seek the resonance and strive to uncover what feels magical out there in the chaos at the edge of the universe. In the immortal words of Trent Reznor: "You and me, We're in this together now." (Nine Inch Nails, 'We're In This Together,' The Fragile, 1999)

Author Note: If your work or interests are aligned with this discussion, feel free to ping me @ lain.mcneill@excessionengineering.com. Otherwise, Happy Valentine's Day everyone. I wish all of you the very best.