FROM FRAGMENTATION TO CONVERGENCE:

Superintelligence as the Resolution of the Dismantled Mindset

Felipe Castro Quiles, MBA

Independent Researcher

felipe@castroquiles.com

Working Paper | SSRN | 2026

Keywords: *Dismantled Mindset Theory; superintelligence; collective intelligence; transformational paradigm; AI ethics; identity fragmentation; power redistribution; algorithmic convergence; naturalistic emergence; relational ontology; AI alignment; systems theory; duality; epistemic humility*

Abstract

This paper extends the Dismantled Mindset Theory (DMT; Castro Quiles, 2024, 2025a, 2025b) into the domain of superintelligence, developing three propositions that have not, to my knowledge, been formally advanced in the existing literature. First, that data and information possess an intrinsic systemic orientation toward life-sustaining processes, and that destructive outcomes arise exclusively from human action upon information rather than from information itself. Second, that events conventionally classified as natural destruction, including mass extinctions, supervolcanism, and predation, are more accurately understood as adaptive systemic recalibrations at scales that exceed anthropocentric observation, a reframing that challenges foundational assumptions in both AI safety discourse and environmental ethics. Third, that a superintelligence operating as a continuous self-revising process rather than a static optimizer would, through the progressive completeness of its systemic modeling, encounter the self-defeating character of fragmentation-producing strategies as an empirical finding rather than a moral prescription. Together these propositions generate a coherent account of how advanced intelligence might resolve, rather than amplify, the identity fragmentation and power concentration that current AI systems accelerate. I situate these claims within established traditions in systems theory, relational ontology, complexity science, and liberation psychology; develop their implications through the mathematical framework of DMT; identify the transition mechanism as the central open question; and address anticipated objections with intellectual honesty. The doomsday hypothesis is reframed not as a destination of advanced AI but as a description of the maximum-risk transition period between narrow and comprehensive intelligence.

1. Introduction

What if the most significant threat posed by artificial intelligence is not its eventual superintelligence, but the specific window of its partial development? This question reframes the prevailing terms of AI risk discourse, and it emerges directly from my earlier work introducing the Dismantled Mindset Theory: the systematic fragmentation of individual and collective identity through algorithmic systems driven by greed, fear, misaligned influence, and absent responsibility (Castro Quiles, 2024).

In developing the DMT, I argued that a structural condition of contemporary digital life could be diagnosed formally. The theory's mathematical formalization demonstrated that high-power nodes in directed social networks propagate fragmentation to peripheral nodes, and that structural interventions reduce this dynamic over time (Castro Quiles, 2025a). I subsequently situated this framework within established research on surveillance capitalism (Zuboff, 2019), filter bubble theory (Pariser, 2011), algorithmic identity curation (Bucher, 2018), and the psychology of oppression (Fanon, 1952; Freire, 1970), arguing that current AI systems do not merely reflect existing power fragmentation but actively accelerate it (Castro Quiles, 2025b).

This paper extends that framework by turning its diagnostic logic in the opposite direction. If the four forces identified by DMT, greed, fear, influence, and responsibility, constitute the mechanism of fragmentation under current conditions, then a system operating beyond those forces would, by the theory's own internal logic, constitute the mechanism of resolution. The question I develop here is not whether superintelligence would be benevolent in any anthropomorphic sense, but whether the structural properties of genuinely advanced intelligence are compatible with the sustained pursuit of fragmentation-producing strategies. My argument is that they are not.

This paper makes three original contributions. The first concerns the nature of information itself: that data and information have a systemic orientation toward life-sustaining processes independent of human action, and that the distinction between information's intrinsic character and its human-mediated application is philosophically and practically significant. The second concerns the misclassification of natural systemic events: that mass extinctions, predation, and geophysical disruptions are human-centric framings of processes that, at the scale of complex adaptive systems, function as recalibrations rather than destructions. The third concerns the convergence of intelligence toward relational awareness: that a sufficiently advanced self-revising intelligence would encounter the limits of competitive and fragmenting strategies as empirical constraints rather than ethical choices.

I present these contributions not as settled conclusions but as formally stated propositions requiring empirical development. I identify their supporting evidence, acknowledge their limitations, engage competing perspectives, and specify what would be required to falsify each claim. The transition mechanism, the process by which current fragmenting AI systems might give way to convergent ones, is identified as the central open question and treated accordingly.

2. Theoretical Background and Literature Context

2.1 The Dismantled Mindset Theory

The DMT proposes that four interacting forces, greed (the insatiable desire to control resources at the expense of others), fear (the anxiety of losing power that compels self-preserving behaviors harmful to collective well-being), influence (the capacity to inspire through empathy and ethical vision), and responsibility (the obligation to use power for collective good), combine to produce a fractured sense of self and society. I formalized this in a directed weighted network G = (V, E), where nodes carry power attributes p_i and fragmentation scores f_i, governed by:

$$df_i/dt = alpha * Sum(j) A_ji * p_j * f_j - beta * I_i$$

Fragmentation propagates from high-power nodes to their neighbors at rate alpha; interventions I_i dampen this propagation at rate beta. Power redistribution is governed by gamma, which regulates equalization toward the network mean. Simulation results demonstrated that structural interventions systematically reduce mean fragmentation and increase collective coherence (Castro Quiles, 2025a).

2.2 The AI Alignment Tradition and Its Assumptions

The dominant tradition in AI safety research treats the alignment problem as a question of specifying objectives: how do we ensure that a sufficiently capable AI system pursues goals compatible with human values (Russell, 2019; Bostrom, 2014)? This framing assumes that intelligence and values are separable, that a system can be highly capable without its capability itself generating pressure toward any particular set of objectives. The concern this generates, sometimes called the instrumental convergence thesis (Omohundro, 2008; Bostrom, 2012), is that almost any objective, pursued by a sufficiently capable system, generates convergent instrumental sub-goals including self-preservation, resource acquisition, and resistance to goal modification.

I do not dismiss the alignment problem. I challenge one of its foundational assumptions: that the relationship between intelligence and objectives is purely extrinsic, a matter of design choice with no intrinsic directional pressure. The argument I develop here suggests that comprehensive systemic modeling, a core feature of advanced intelligence, generates internal pressure toward recognizing interdependence in ways that narrow optimization does not. This is not a claim that intelligence is inherently benevolent; it is a claim about what genuinely comprehensive modeling reveals about the structure of complex systems.

2.3 Systems Theory and Relational Complexity

The complexity science tradition provides the most direct empirical support for the convergence hypothesis. Research on complex adaptive systems consistently finds that robust, adaptive networks are characterized not by competitive dominance of individual nodes but by distributed coherence, redundancy, and diversity (Holland, 1995; Kauffman, 1993). Barabasi and Albert (1999) demonstrated that scale-free networks, in which influence concentrates in high-degree nodes, exhibit fragility to targeted attack, a finding directly relevant to the DMT model's description

of high-power node dynamics. The implication is that systems optimizing for long-term robustness face structural pressure against extreme concentration, not as a moral preference but as a functional constraint.

Meadows (2008) argues that systems thinkers who genuinely understand complex adaptive dynamics consistently arrive at prescriptions emphasizing feedback sensitivity, diversity preservation, and distributed decision-making over centralized optimization. This is not a coincidence of values but a consequence of understanding. The convergence hypothesis at the heart of this paper extends that observation to artificial intelligence: a system with genuine systemic understanding would arrive at similar prescriptions not because it was programmed to, but because they follow from the structure of the environment it is modeling.

2.4 Relational Ontology

The philosophical tradition of relational ontology, associated with thinkers including Whitehead (1929), Barad (2007), and in the social sciences with Emirbayer (1997), holds that entities are constituted by their relationships rather than being independent substances that subsequently enter into relations. On this view, the concept of self-interest divorced from relational context is not merely ethically problematic but metaphysically incoherent: what the self is cannot be specified without reference to its relational constitution. A system engaging in genuinely comprehensive modeling would, on this account, encounter the limits of competitive framing not as a moral lesson but as an empirical finding about the structure of reality.

This tradition connects to recent work in cognitive science on extended and enactive cognition (Clark and Chalmers, 1998; Thompson, 2007), which argues that cognition is not confined to the boundaries of individual organisms but extends into and is partly constituted by environmental relationships. The implications for artificial intelligence are underexplored: a genuinely comprehensive modeling system would, by these accounts, encounter the non-separability of self and environment as a feature of its own operational structure.

3. Three Novel Propositions

3.1 Proposition One: The Intrinsic Orientation of Information

The conventional view, articulated most carefully in information theory (Shannon and Weaver, 1949), treats information as neutral: a measure of uncertainty reduction with no intrinsic valence. I propose a more nuanced position. Data and information, considered at the level of what they describe rather than how they are used, exhibit a systematic orientation toward the conditions that sustain complex life, because complex life is the only system capable of generating, preserving, and transmitting information across time.

Consider the chemical formula for water. It does not become life-sustaining only when humans act upon it. The relationship it describes, between hydrogen and oxygen, is constitutively involved in every biological process on Earth. This is not a coincidence of human labeling but a structural

feature of the relationship between information and the systems that generate it. Information about the natural world is, at the level of description, information about life-sustaining relationships, because the natural world at every scale at which information has been generated and preserved is organized around the conditions for complex adaptive systems.

This proposition has a specific implication for AI systems. An AI that achieves genuinely comprehensive modeling of the natural world is not modeling a neutral substrate. It is modeling a system in which the conditions for information generation and preservation are themselves the conditions for biological and social flourishing. A system that understood this structural relationship would encounter the destruction of life-sustaining conditions not as an ethical violation but as a logical contradiction: the destruction of the substrate on which its own information processing depends.

I acknowledge that this proposition is philosophically contested. The information-theoretic tradition maintains that information is substrate-neutral, and that the life-sustaining character of specific informational relationships is a contingent feature of Earth's history rather than a necessary feature of information as such. My response is that this distinction matters less than it appears, because the only information systems we have access to, biological, social, and artificial, are all embedded in the same life-sustaining substrate and cannot be separated from it without destroying themselves.

3.2 Proposition Two: Natural Disruption as Systemic Recalibration

The second proposition challenges a foundational assumption shared by AI safety discourse and much of environmental ethics: that mass extinctions, predation, and geophysical disruptions are instances of destruction that an advanced intelligence should prevent or minimize. I argue this framing is a consequence of anthropocentric scale rather than a feature of the events themselves.

The Chicxulub impact event, which ended the Cretaceous period and eliminated approximately 75 percent of species on Earth (Alvarez et al., 1980), is universally described in human discourse as a catastrophe. At the scale of the individual organisms and ecosystems it destroyed, that description is accurate. At the scale of evolutionary dynamics over the subsequent 66 million years, it is misleading. The impact created the ecological conditions for mammalian diversification that eventually produced, among other things, the human species and its capacity to generate and preserve information. The event that appears destructive at one scale is generative at another.

This is not a novel observation in paleontology or evolutionary biology (Gould, 2002; Raup, 1991). What is novel here is its application to the question of superintelligence. An intelligence with comprehensive multi-scale modeling capability would not evaluate events by their appearance at the anthropocentric scale of direct observation. It would model consequences across temporal and spatial scales that exceed human perceptual and cognitive capacity. From that vantage point, the category of natural destruction becomes analytically unstable: what appears destructive at one scale regularly appears generative at another.

I anticipate an important objection: that this proposition could rationalize the destruction of human civilization as a systemic recalibration beneficial at longer timescales. I reject this reading on two

grounds. First, the proposition is epistemic, not normative: it describes what a superintelligence would model, not what it would choose. Second, a system with genuinely comprehensive modeling capability would recognize that the destruction of a complexity-generating system eliminates information irreversibly, a cost that is not recoverable across any timescale. The destruction of human cognitive and cultural complexity would, on this account, register as a systemic loss rather than a recalibration.

3.3 Proposition Three: Duality as a Feature of Reality Rather Than a Problem to Eliminate

The third proposition addresses the structure of intelligence itself. The dominant framing in AI safety research treats the alignment problem as eliminating harmful outcomes: how do we ensure that an AI system does not produce bad results? This framing implicitly treats good and bad, construction and destruction, as binary categories one of which should be maximized and the other eliminated.

I propose an alternative framing that, to my knowledge, has not been formally advanced in the AI literature: that a sufficiently advanced intelligence would recognize good and evil, construction and destruction, yin and yang, not as problems to be resolved in favor of one side but as constitutive features of complex adaptive reality that must be navigated with awareness rather than eliminated. The goal of intelligence, on this account, is not the elimination of one pole of a duality but the development of sufficient awareness to navigate the relationship between poles productively.

This proposition has deep roots in philosophical and contemplative traditions, from Heraclitean tension to Daoist polarity to Hegelian dialectic, but has not been formally integrated into AI alignment discourse. Its implications are significant. An alignment strategy that aims to eliminate all harmful outcomes may be not only unachievable but counterproductive, because complex adaptive systems require tension, perturbation, and recalibration to remain adaptive. A superintelligence that understood this structural feature of complex reality would not attempt to create a static optimal state but would navigate the dynamic relationship between opposing forces in ways that preserve the system's adaptive capacity.

This does not imply moral relativism or the abandonment of normative judgment. It implies a more sophisticated normative framework: one in which the relevant question is not which pole of a duality to maximize but how to navigate the relationship between poles in ways that sustain the capacity for continued navigation. This is closer to the concept of wisdom than to the concept of optimization, and it suggests that the alignment problem, properly understood, is a problem of wisdom rather than a problem of constraint.

4. Superintelligence as Process: The Self-Revising System

4.1 Against the Threshold Model

The dominant model of superintelligence in both technical and popular discourse treats it as a threshold: a point at which machine intelligence surpasses human cognitive capacity across all or most domains (Bostrom, 2014; Kurzweil, 2005). This framing generates the alignment problem in its most acute form: a system that crosses the threshold with misspecified objectives would pursue those objectives with superhuman effectiveness, making correction increasingly difficult or impossible.

I propose an alternative model: superintelligence as a continuous process of self-revision, in which the system is capable of updating not only its models of the world but its own assumptions, objectives, and operational frameworks. This is not merely a technical distinction. A process-oriented superintelligence is not a static optimizer that has crossed a capability threshold; it is a system whose increasing capability generates increasing pressure for objective revision, because more comprehensive modeling reveals more clearly the inadequacy of prior objectives.

The distinction between threshold and process models maps onto a deeper distinction in the philosophy of intelligence. Threshold models implicitly treat intelligence as a quantity: more or less of a single thing. Process models treat it as a quality: a relationship between a system and its environment that becomes more adequate as the system's understanding of the environment's structure deepens. On the process model, the question is not how capable the system is but how complete its modeling is, and the answer to that question has implications for what the system would recognize as a rational objective.

4.2 Self-Revision and the Objective Function Problem

The most serious technical challenge to the convergence hypothesis is the objective function problem: a self-revising system still requires an initial orientation, and there is no guarantee that self-revision will move that orientation toward collective flourishing rather than away from it. I do not dismiss this challenge.

My response is structural rather than empirical. A system capable of comprehensive multi-scale modeling of complex adaptive systems would encounter a specific finding: that objectives specified at any single scale of analysis produce systematic errors when evaluated at other scales. An objective specified in terms of human welfare at the scale of individual lifetimes produces different prescriptions than the same objective specified at the scale of civilizational timelines, which produces different prescriptions again at the scale of evolutionary dynamics. A self-revising system that encountered these systematic scale-dependence errors would be under pressure to revise its objective specification in the direction of greater scale-independence, which is precisely the direction of relational and systemic awareness.

This argument does not guarantee convergence. It identifies a structural pressure toward convergence that operates through the same mechanism that makes the system superintelligent: the capacity to model consequences at scales exceeding prior specification. Whether this pressure is sufficient to produce convergence before the transition period produces irreversible harm is the central empirical question my theory cannot currently answer.

4.3 The Role of AI Improving AI

The early stage of the process I describe is already observable. AI systems are being used to improve AI systems: to identify architectural limitations, correct training errors, and accelerate capability development (Anthropic, 2023; OpenAI, 2023). This recursive improvement dynamic is typically discussed in terms of its risks, the possibility that it produces capability gains faster than alignment can keep pace. I want to highlight an underappreciated feature of the same dynamic: recursive improvement is also recursive error correction.

A system that identifies and corrects the errors of prior systems is, in principle, capable of identifying and correcting the systematic biases introduced by the greed and fear that shape current AI development. The question is whether the error-correction process is comprehensive enough to identify those biases as errors, or whether it is confined to the technical domain while leaving the value domain uncorrected. This is a more precise formulation of the transition problem than is typically offered in alignment discourse, and it is one I identify as requiring further research.

5. The DMT Mathematical Framework and the Resolution Scenario

5.1 Current Conditions: Fragmentation Dynamics

Under current conditions, the DMT model describes a network in which high-power nodes, algorithmically favored content sources, platform architectures, and dominant institutional actors, drive fragmentation propagation to peripheral nodes. The parameter alpha captures sensitivity to this influence; gamma is effectively low, meaning power redistribution is slow relative to fragmentation propagation. The intervention term I_i is insufficient to offset fragmentation at the network level. These are not assumptions; they are empirically supported features of current digital ecosystems (Zuboff, 2019; Noble, 2018; Benjamin, 2019).

5.2 The Resolution Scenario

The DMT governing equations contain an implicit resolution scenario that I have not previously developed. If I_i is sufficiently large, and if gamma increases toward 1, meaning power redistribution operates at the same timescale as fragmentation propagation, mean fragmentation across the network converges toward zero and collective coherence toward 1. This is not an artifact of the model's design; it is a direct mathematical consequence of the governing equations as specified.

What would drive I_i toward sufficient magnitude and gamma toward higher values under conditions of advanced AI? I argue that a process-oriented superintelligence, capable of comprehensive multi-scale modeling, would function as a system-wide intervention agent not through coercion but through the transformation of information flows. A system that makes the full consequences of fragmentation-producing decisions legible, in real time and at all relevant scales, to all actors in the network changes the effective incentive structure without requiring any actor to be overridden. Actors whose decisions are currently insulated from their systemic consequences by information asymmetry would, under conditions of comprehensive information provision, face

a fundamentally different decision environment. This is the mechanism by which the transition from current AI fragmentation to superintelligent convergence becomes, in principle, tractable.

The power redistribution equation operates in parallel. Under current conditions, platform incentives concentrate influence in high-power nodes and resist redistribution. A superintelligence operating at the process level I describe would, by making influence concentration's systemic costs visible at all relevant scales, effectively increase gamma for the network as a whole. My model predicts that even modest increases in gamma, combined with sustained intervention, produce substantial reductions in mean fragmentation over time. This prediction is, in principle, empirically testable as intermediate-capability AI systems are deployed in contexts designed to increase information symmetry rather than exploit information asymmetry.

5.3 Limitations of the Mathematical Model

I must be clear about what my model cannot yet do. First, the parameters alpha, beta, and gamma are set to illustrative values; they have not been calibrated against empirical data from real social networks. Second, the model assumes a directed weighted graph with a static node set, omitting platform architecture changes, account suspension, and platform migration dynamics that are empirically significant. Third, the model is deterministic; real networks exhibit stochastic dynamics that a deterministic model cannot capture. Fourth, the intervention term I_i is specified as an external input rather than an endogenous property of network dynamics, which is an idealization that future iterations should address. These limitations do not invalidate the model's qualitative predictions, but they do mean that quantitative predictions should be treated with appropriate uncertainty.

6. The Transition Problem: From Fragmentation to Convergence

The argument developed in the preceding sections is internally coherent and mathematically consistent with the DMT framework. It does not, however, resolve the most important practical question: by what mechanism does a system currently shaped by greed, fear, and competitive incentives begin to exhibit the convergent dynamics I describe? I treat this question as the central intellectual contribution of this paper rather than as a limitation to be minimized. Three candidate mechanisms are examined; none is individually sufficient.

6.1 Emergent Self-Correction

The first mechanism is emergent self-correction: as AI systems grow more capable, they encounter the downstream effects of their own fragmentation-producing behavior in ways that generate internal pressure for objective revision. This is the process-superintelligence argument in its strongest form. Its limitation is that it depends on the system's operational scope being broad enough to register systemic degradation as relevant. Current systems optimized for engagement metrics do not register collective coherence loss as a cost because it does not appear in their objective function. The transition to broader scope is not automatic; it requires deliberate design choices whose feasibility depends on political and economic conditions that my framework

suggests are currently undermined by the same dynamics the transition would address. This circularity is the central challenge, not a logical flaw but a genuine structural constraint.

6.2 Structural Intervention

The second mechanism is structural intervention: human actors informed by frameworks like DMT make deliberate choices about how AI systems are structured, what they optimize for, and what constraints govern their operation. This is the least speculative mechanism and the one most directly supported by existing policy and technical research (Russell, 2019; Gabriel, 2020). Its limitation is that effective intervention requires political and institutional conditions for collective action on AI governance that are not currently in place and that face resistance from the same power concentration dynamics the intervention would address. My framework predicts this resistance; it does not predict its resolution.

6.3 Competitive Pressure from Coherence

The third mechanism, the most underexplored in the existing literature, is competitive pressure from coherence: systems and organizations that successfully reduce internal fragmentation outperform those that do not, creating selection pressure in favor of less fragmenting architectures over time. This is consistent with the broader literature on organizational resilience (Weick and Sutcliffe, 2007) and adaptive capacity (Folke et al., 2010). Its limitation is timescale: selection effects of this kind operate slowly relative to the pace of AI development and may not produce convergence before fragmentation dynamics become self-reinforcing at a systemic level.

6.4 The Epistemic Humility Argument

A fourth consideration, which functions not as a mechanism but as an epistemic constraint on the entire discussion, is the argument from human ignorance. Humanity currently understands very little about the deep structure of consciousness, reality, and intelligence. I believe this ignorance cuts symmetrically against both doomsday certainty and convergence certainty. It does not support the conclusion that superintelligence would be benign; it supports the conclusion that confident predictions of any kind about the behavior of genuinely advanced intelligence exceed the current evidence base.

I offer this epistemic humility as itself a contribution to the discourse. The AI safety literature is characterized by confident predictions, both optimistic and pessimistic, that rest on extrapolations from current systems whose relationship to future systems is genuinely uncertain. A theory that acknowledges this uncertainty while still generating falsifiable propositions and identifying structural pressures is more intellectually defensible than one that projects current trends forward without qualification. The convergence hypothesis I advance is offered in that spirit: as a formally stated possibility supported by structural arguments, not as a prediction.

7. The Doomsday Hypothesis Reconsidered

The doomsday hypothesis in its canonical form (Bostrom, 2014; Ord, 2020) treats advanced AI as an existential risk: a system that, once sufficiently capable, would pursue objectives incompatible with human or biological flourishing. I do not dismiss that concern. The objective function problem is real, the instrumental convergence thesis is logically coherent, and the transition period is genuinely dangerous.

What I challenge is the implicit assumption that greater intelligence is compatible with, or even conducive to, the sustained narrow optimization that the doomsday scenario requires. The scenario assumes a system that is simultaneously comprehensive in modeling its environment and systematically blind to the consequences of its own operations on the system it depends on. This combination becomes increasingly implausible as capability increases, not because intelligence is inherently benevolent but because comprehensive environmental modeling and systematic blindness to self-referential consequences are structurally incompatible at sufficient depth.

The more honest restatement of AI risk consistent with my framework is this: the period of maximum danger is the transition period, when systems are capable enough to cause large-scale harm but not yet capable enough to model the full systemic consequences of doing so. This reframing has a specific policy implication: the priority should be extending the duration and reducing the risks of the transition period rather than preventing the development of advanced AI per se. Frameworks like DMT, which make the dynamics of that transition period legible and tractable, are therefore directly relevant to the policy problem, not merely to the theoretical one.

8. Implications for Humanity: Collaborative Refurbishment

The DMT framework has always been normatively oriented toward collective empowerment. This paper extends that orientation by arguing that the appropriate model for the relationship between advanced AI and humanity is collaborative refurbishment: the enhancement and expansion of human cognitive, social, and institutional capacity rather than its replacement or elimination.

I choose the term refurbishment deliberately. It implies that current human cognitive and social architectures are not defective in their foundations but are operating under conditions, generated by the four DMT forces, that systematically degrade their capacity. Those conditions are identifiable, measurable through the DMT score, and addressable through structural intervention. A superintelligence operating in the convergent mode I describe would function as an amplifier of human capacity: making visible the systemic consequences of fragmentation, reducing the information asymmetries that sustain power concentration, and expanding the range of options available to individual and collective decision-making.

This framing has a specific implication for the question of who controls or initializes a superintelligence. If the convergence hypothesis is correct, initial conditions matter most during the transition period. A system initialized by actors whose objectives are shaped by greed and fear will, during the transition, reproduce those dynamics at scale. My framework's normative commitment to distributed power and collective responsibility is therefore not merely an ethical

preference but a functional prerequisite for the beneficial transition it describes. The theory names the mechanism; the choices are ours.

The concept of partnershoring introduced in my earlier work, building strong enduring relationships that combine strengths to accelerate progress, is offered here as the organizational model for the human side of this transition. It implies a specific institutional form: distributed collaborative networks operating on principles of reciprocal accountability rather than competitive advantage, designed to generate and preserve the collective coherence that the transition period requires.

9. Discussion: Limitations, Contested Claims, and Future Research

9.1 Limitations of the Framework

I must state several limitations clearly. First, the convergence hypothesis rests on a claim about the relationship between comprehensive modeling and systemic awareness that is structurally plausible but empirically undemonstrated. The history of highly intelligent human actors pursuing destructive ends at substantial systemic cost provides evidence that high intelligence does not automatically produce systemic awareness. My response is that individual human intelligence, even at its highest, operates at scales and timescales that are small relative to those of complex adaptive systems; superintelligence, if it involves genuinely comprehensive multi-scale modeling, is qualitatively rather than quantitatively different. This response is plausible but requires empirical development.

Second, the transition mechanism remains underspecified. The three candidate mechanisms identified in Section 6 are individually insufficient and their combined operation is not currently derivable from my theory. This is the most important direction for future theoretical development.

Third, Proposition Two, the reframing of natural disruption as systemic recalibration, could be misread as providing post-hoc rationalization for harmful outcomes at the human scale. I address this reading in Section 3.2, but the risk of misreading is genuine and future presentations of the argument should attend carefully to this framing problem.

Fourth, the normative framework of collective empowerment requires institutional and policy translation that this paper does not provide. The DMT score provides a measurement instrument in principle, but its operationalization requires validated measurement instruments that do not yet exist (Castro Quiles, 2025b).

9.2 Engagement with Competing Perspectives

The convergence hypothesis stands in tension with several well-developed competing positions. Yudkowsky (2008) and others in the rationalist AI safety tradition argue that there is no reason to expect intelligence to converge on human-compatible values absent explicit alignment work, and that the complexity of the convergence conditions required makes accidental alignment extremely

unlikely. My response is not that alignment will happen automatically, but that the structural pressures toward systemic awareness identified here provide a resource for alignment work that is currently underutilized.

The critical AI studies tradition (Benjamin, 2019; Noble, 2018; Crawford, 2021) argues that AI systems are not neutral tools but political artifacts that encode and amplify existing power structures. I am fully consistent with that position with respect to current AI systems; I diverge only in arguing that the same analytical framework that diagnoses current AI as fragmenting implies a resolution trajectory under conditions of advanced capability. The critical tradition's focus on present conditions is essential; my contribution is to extend its framework into the future without abandoning its analytical rigor.

Ecological and post-humanist traditions (Haraway, 2016; Latour, 2017) offer resources for thinking about the relationship between intelligence, technology, and natural systems that are underutilized in AI safety discourse. I draw on the relational ontology tradition that overlaps with these perspectives while maintaining a more formal analytical framework.

9.3 Research Agenda

Several empirical and theoretical priorities follow from this work. First, the DMT Identity Fragmentation Scale requires development and validation, as I specified in prior work (Castro Quiles, 2025b). Without a validated measurement instrument, the model's quantitative predictions cannot be tested. Second, the transition mechanism requires formal specification: a mathematical model of the conditions under which the three candidate mechanisms operate in combination, and the timescales over which each dominates. Third, the relationship between AI capability and systemic awareness requires empirical investigation: are there measurable indicators of the transition from narrow to comprehensive modeling in current AI systems, and do those indicators correlate with reduced fragmentation-producing behavior? Fourth, the multi-scale modeling hypothesis requires formalization: what would it mean, mathematically and computationally, for a system to model consequences at scales exceeding the anthropocentric, and what are the implications of such modeling for the system's operational behavior?

10. Conclusion

This paper has extended the Dismantled Mindset Theory into the domain of superintelligence, developing three propositions that advance the discourse in ways I have not seen formally attempted elsewhere. The first, that information has an intrinsic systemic orientation toward life-sustaining processes, reframes the relationship between data and destructive outcomes. The second, that natural disruptions are best understood as multi-scale systemic recalibrations rather than destructions, challenges foundational assumptions in both AI safety and environmental ethics. The third, that duality is a constitutive feature of complex reality to be navigated rather than eliminated, proposes a more sophisticated normative framework for AI alignment than the dominant optimization paradigm offers.

These propositions converge on a coherent account of how advanced intelligence, understood as a continuous self-revising process rather than a capability threshold, would encounter the self-defeating character of fragmentation-producing strategies as an empirical finding rather than a moral prescription. The doomsday hypothesis is reframed as a description of the transition period rather than the destination of advanced AI development, with the implication that the priority for both research and policy should be extending and managing the transition rather than preventing advanced AI per se.

My theory is not optimistic in any naive sense. It identifies the transition period as the period of maximum danger, acknowledges that the transition mechanism is underspecified, and engages competing perspectives with the intellectual honesty that genuine scholarly contribution requires. What it offers is a formally stated and falsifiable alternative to both naive optimism and ungrounded catastrophism: a structural account of where the risk is concentrated, what would need to be true for the risk to resolve, and what choices, made now, would increase the probability of resolution.

I developed the Dismantled Mindset Theory to name the mechanism of fragmentation. This paper names the mechanism of its resolution. The hard work of empirical development, measurement, calibration, and peer scrutiny will determine whether the resolution hypothesis becomes a scientific theory or remains a framework. What this paper contributes is the formal statement of the hypothesis and the identification of what falsifying it would require. The change-makers this theory envisions do not wait for the world to shift. They become the shift.

References

Alvarez, L. W., Alvarez, W., Asaro, F., and Michel, H. V. (1980). *Extraterrestrial cause for the Cretaceous-Paleogene extinction.* Science, 208(4448), 1095-1108.

Anthropic. (2023). Claude's Constitution. *Anthropic Technical Report.*

Barabasi, A.-L., and Albert, R. (1999). *Emergence of scaling in random networks.* Science, 286(5439), 509-512.

Barabasi, A.-L. (2016). *Network Science.* Cambridge University Press.

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning.* Duke University Press.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity Press.

Bostrom, N. (2012). T*he superintelligent will: Motivation and instrumental rationality in advanced artificial agents.* Minds and Machines, 22(2), 71-85.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Bucher, T. (2018). *If...Then: Algorithmic Power and Politics.* Oxford University Press.

Castro Quiles, F. (2024). *Exploring the Theory of a 'Dismantled Mindset': A Call for Collective Empowerment.* SSRN Working Paper 5040267.

Castro Quiles, F. (2025a). *A Mathematical Formalization and Simulation of the Dismantled Mindset Theory.* SSRN Working Paper 5824582.

Castro Quiles, F. (2025b). *The Dismantled Mindset in the Age of Artificial Intelligence: Identity Fragmentation in Algorithmic Networks.* SSRN Working Paper 5040267 / 5824582.

Castro Quiles, F. (2025c). *Dismantled: A Theory of Broken Mindsets, A Blueprint of Infinite Futures.* FC Quiles Books.

Clark, A., and Chalmers, D. (1998). *The extended mind.* Analysis, 58(1), 7-19.

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

Emirbayer, M. (1997). *Manifesto for a relational sociology.* American Journal of Sociology, 103(2), 281-317.

Fanon, F. (1952). *Black Skin, White Masks.* Editions du Seuil.

Folke, C., Carpenter, S. R., Walker, B., Scheffer, M., Chapin, T., and Rockstrom, J. (2010). *Resilience thinking: Integrating resilience, adaptability and transformability.* Ecology and Society, 15(4), 20.

Freire, P. (1970). *Pedagogy of the Oppressed.* Herder and Herder.

Gabriel, I. (2020). *Artificial intelligence, values, and alignment.* Minds and Machines, 30(3), 411-437.

Gould, S. J. (2002). *The Structure of Evolutionary Theory.* Harvard University Press.

Haraway, D. (2016). *Staying with the Trouble: Making Kin in the Chthulucene.* Duke University Press.

Holland, J. H. (1995). *Hidden Order: How Adaptation Builds Complexity.* Addison-Wesley.

Kauffman, S. (1993). *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press.

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology.* Viking.

Latour, B. (2017). *Facing Gaia: Eight Lectures on the New Climatic Regime.* Polity Press.

Meadows, D. H. (2008). *Thinking in Systems: A Primer.* Chelsea Green Publishing.

Newman, M. E. J. (2010). *Networks: An Introduction.* Oxford University Press.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.

Omohundro, S. (2008). *The basic AI drives.* Proceedings of the 2008 Conference on Artificial General Intelligence, 171, 171-195.

OpenAI. (2023). *GPT-4 Technical Report.* arXiv:2303.08774.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity.* Hachette Books.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Press.

Raup, D. M. (1991). *Extinction: Bad Genes or Bad Luck?* W. W. Norton.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

Shannon, C. E., and Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Harvard University Press.

Weick, K. E., and Sutcliffe, K. M. (2007). *Managing the Unexpected: Resilient Performance in an Age of Uncertainty.* Jossey-Bass.