The Self as a Compression Artifact That Refuses to Die

Abstract

The subjective self is typically treated as a fundamental feature of consciousness or a metaphysical entity. This paper reframes selfhood as an information-theoretic problem: the self emerges as a lossy compression of internal processes and persists because removing it destabilizes prediction and control. We model the self as a compressed latent variable S that encodes high-dimensional internal states X into a lower-dimensional representation. Using rate-distortion theory, we show that self-consistency is rewarded over self-accuracy because consistent representations enable better prediction and control, even when they sacrifice fidelity to ground truth. We analyze identity defense, rationalization, and ego threat as compression protection mechanisms that prevent costly recompression. We compare human and artificial self-models, showing both exhibit similar compression dynamics despite different constraints. Finally, we explain why self-erasure (ego death) feels both terrifying and liberating: decompression eliminates the compressed representation but forces the system to process raw, high-entropy information. The analysis treats selfhood as a computational structure, avoiding metaphysics while providing testable predictions about identity maintenance and defense mechanisms.

1 Introduction

The question of what the self is has occupied philosophers for millennia, typically framed in terms of consciousness, personal identity, or metaphysical substance. This paper takes a different approach: we treat the self as an *information structure*—a compressed representation of internal processes that serves computational functions. Rather than asking what the self *is* in some deep metaphysical sense, we ask: *what computational problem does the self solve, and why does it persist even when it conflicts with accuracy?*

This work builds on and formalizes several existing frameworks. Dennett's "center of narrative gravity" [7] treats the self as a narrative construct, while Metzinger's self-model theory [6] frames selfhood as a representational structure. Friston's free-energy principle [8] emphasizes prediction error minimization as a core computational principle. This paper synthesizes these perspectives through an explicit information-theoretic formalization: we model the self as a compressed latent variable using rate-distortion theory, showing how consistency is rewarded over accuracy, and analyzing the computational costs of recompression. What is genuinely new here is the explicit rate-distortion framing, the formalization of the consistency-accuracy tradeoff, and the analysis of recompression costs as drivers of defensive behaviors. This is a formalization and synthesis, not a replacement of existing theories.

The central thesis is that the subjective self emerges as a lossy compression of internal processes and persists because removing it destabilizes prediction and control. This reframing shifts the question from metaphysics to engineering: the self is not a thing that exists independently, but a compression artifact that serves functional purposes. Like JPEG compression discards high-frequency details to reduce file size, the self discards contradictory information, uncertainty, and high-dimensional details to create a manageable representation.

This perspective explains several puzzling features of selfhood that are difficult to account for in traditional frameworks:

- **Why self-consistency trumps accuracy**: The self maintains coherent narratives even when they conflict with evidence, because consistency enables prediction while accuracy requires expensive recompression.
- **Why identity defense is so strong**: Threats to the self-model trigger defensive responses because recompression is computationally costly.
- **Why rationalization is ubiquitous**: Post-hoc compression of inconsistent actions is cheaper than maintaining accurate but contradictory representations.
- **Why ego death is terrifying and liberating**: Removing the compressed representation eliminates maintenance costs but forces processing of raw, high-entropy information.

The paper proceeds as follows. Section 2 models the self as a compressed latent variable using information theory and rate-distortion theory. Section 3 explains why self-consistency is rewarded over self-accuracy through prediction error minimization and control theory. Section 4 analyzes identity defense, rationalization, and ego threat as compression protection mechanisms. Section 5 compares human and artificial self-models, showing similar compression dynamics. Section 6 explains the paradoxical nature of ego death as decompression. Section 7 discusses implications and predictions. Section 8 concludes.

Throughout, we maintain a computational perspective: the self is an information structure, not a metaphysical entity. This allows us to make testable predictions and avoid the circular reasoning that plagues traditional approaches to selfhood.

2 The Self as Compressed Latent Variable

2.1 Information-Theoretic Formulation

Consider a system with internal processes $X = \{X_1, X_2, \ldots, X_n\}$ representing beliefs, memories, goals, emotional states, sensory inputs, and other high-dimensional information. The full state space has entropy $H(X)$, which may be extremely large. The self S is a compressed representation of X:

$$S = f(X_1, X_2, \ldots, X_n) \tag{1}$$

where f is a compression function that maps high-dimensional X to lower-dimensional S. The compression is lossy: $I(S; X) < H(X)$, meaning the self representation contains less information than the full internal state.

The compression ratio is:

$$R = \frac{H(S)}{H(X)} \tag{2}$$

where $R < 1$ indicates compression. Typical self-models have $R \ll 1$: a person's identity can be summarized in a few sentences, while their full internal state over a lifetime would require petabytes of information.

2.2 Rate-Distortion Theory

The self can be formalized using rate-distortion theory [1]. We seek a compressed representation S that minimizes:

$$L = D(X, S) + \lambda R(S) \tag{3}$$

where:

- $D(X, S)$ is the distortion function measuring how much information is lost in compression

- $R(S)$ is the rate (compression cost) measured in bits
- λ is a Lagrange multiplier controlling the tradeoff between distortion and rate

The optimal self-model is:

$$S^* = \arg\min_{S'}[D(X, S') + \lambda R(S')] \tag{4}$$

This formulation makes explicit what gets discarded in compression. The distortion function $D(X, S)$ measures the information-theoretic distance between the full state X and the compressed self S. High-distortion compression discards more information but achieves lower rate (smaller S).

2.3 Distortion Function Specification

The distortion function $D(X, S)$ is not unique—different systems may optimize different distortion metrics while preserving the same compression dynamics. This is a strength of the framework: it captures general compression principles without requiring a specific distortion measure. Plausible distortion metrics include:

Predictive error: $D_{\text{pred}}(X, S) = \mathbb{E}[(Y - \hat{Y}(S))^2]$, where Y are outcomes and $\hat{Y}(S)$ are predictions based on S. This measures how well S enables prediction of future states.

Control error: $D_{\text{control}}(X, S) = \mathbb{E}[(Y_{\text{desired}} - Y_{\text{actual}}(S))^2]$, measuring how well S enables achieving desired outcomes through action selection.

Social signaling loss: $D_{\text{social}}(X, S) = \text{KL}(P(\text{others' model of self}|X)||P(\text{others' model of self}|S))$, measuring how well S enables accurate social communication and coordination.

Narrative incoherence: $D_{\text{narrative}}(X, S) = \text{inconsistency}(S) + \text{gap\_penalty}(X, S)$, measuring contradictions within S and gaps between S and important aspects of X.

Different systems may weight these differently or use combinations. A social species might weight D_{social} heavily, while a solitary agent might emphasize D_{pred} and D_{control}. The key insight is that regardless of the specific distortion function, the same compression dynamics emerge: consistency is rewarded, recompression is costly, and defensive mechanisms protect the compressed representation. This generality is a feature, not a bug—it shows that compression dynamics are fundamental properties of self-modeling systems, not artifacts of specific distortion metrics.

2.4 What Gets Discarded

The compression process discards several types of information to maintain a manageable self-representation:

High-frequency details: Moment-to-moment fluctuations in mood, attention, and physiological states are averaged out. The self-model represents "I am generally optimistic" rather than tracking every micro-fluctuation in affect.

Contradictions: Conflicting beliefs, inconsistent memories, and contradictory goals are resolved or suppressed. The self-model maintains "I am honest" even when specific memories of dishonesty exist, by either forgetting them or rationalizing them.

Uncertainty: Ambiguity about one's traits, abilities, or past is reduced to definite statements. "I am uncertain about my intelligence" becomes "I am intelligent" or "I am not intelligent," but rarely remains as uncertainty in the self-model.

Context-dependence: The self-model represents stable traits across contexts, discarding context-specific variations. "I am extroverted" is maintained even though one might be introverted in specific situations.

Temporal details: Specific memories are compressed into narrative summaries. The self-model contains "I had a difficult childhood" rather than every specific event.

This compression is not arbitrary—it optimizes for functions that require stable, low-dimensional representations: prediction, control, social signaling, and planning.

2.5 Latent Variable Structure

The self S acts as a latent variable in a generative model. The system uses S to predict future states Y:

$$P(Y|X) \approx P(Y|S) \tag{5}$$

where the approximation holds because S captures the relevant information in X for prediction. The self-model enables efficient prediction by working with the compressed representation rather than the full state space.

Similarly, the self enables control by providing a stable reference for action selection:

$$a^* = \arg\max_a U(a, S) \tag{6}$$

where actions are selected to optimize utility U given the self-model S. Without S, the system would need to optimize over the full state space X, which is computationally intractable.

3 Why Self-Consistency Over Self-Accuracy

3.1 Prediction Error Minimization

The self-model is optimized for prediction, not accuracy. Consider the prediction error:

$$L_{\text{pred}} = \mathbb{E}[(Y - \hat{Y}(S))^2] \tag{7}$$

where Y is the actual outcome and $\hat{Y}(S)$ is the prediction based on self-model S. A consistent self-model enables better predictions even when it is inaccurate, because consistency provides a stable basis for prediction.

To see this, consider two self-models:

- $S_{\text{consistent}}$: "I am intelligent and capable" (consistent but may be inaccurate)
- S_{accurate}: "I am sometimes intelligent, sometimes not, depending on context and mood" (accurate but inconsistent)

The consistent model enables predictions: "I will perform well on this task because I am intelligent." The accurate model provides no clear prediction: "I might perform well or poorly, depending on factors I cannot fully specify."

The consistent model may have higher prediction error in some cases, but it has *lower variance* in predictions, which is often more valuable for control. This is the bias-variance tradeoff applied to self-models: consistency introduces bias but reduces variance.

3.2 Control Theory Perspective

From a control theory perspective, a stable self-representation is necessary for effective action selection. Consider the control objective:

$$L_{\text{control}} = \mathbb{E}[(Y_{\text{desired}} - Y_{\text{actual}})^2] + \alpha \cdot \text{inconsistency}(S) \tag{8}$$

where the inconsistency term penalizes self-models that change rapidly or contain contradictions. The system optimizes:

$$S^* = \arg\min_S [L_{\text{pred}} + L_{\text{control}}] \tag{9}$$

The inconsistency penalty $\alpha \cdot \text{inconsistency}(S)$ is typically large, meaning the system strongly prefers consistent self-models even when they sacrifice accuracy. This explains why people maintain stable identities even when evidence contradicts them.

3.3 The Consistency-Accuracy Tradeoff

The full objective function is:

$$L = \mathbb{E}[(Y - \hat{Y}(S))^2] + \alpha \cdot \text{inconsistency}(S) + \beta \cdot \text{inaccuracy}(S) \tag{10}$$

where:

- α is the consistency weight (typically large)
- β is the accuracy weight (typically small)
- $\text{inconsistency}(S)$ measures contradictions and temporal instability in S
- $\text{inaccuracy}(S)$ measures deviation from ground truth

In practice, $\alpha \gg \beta$: consistency is rewarded far more than accuracy. This is because:

1. Consistent models enable stable predictions and control
2. Recompression to achieve accuracy is computationally expensive
3. Accuracy requires maintaining high-dimensional, context-dependent information
4. Consistency can be maintained with low-dimensional, stable representations

3.4 Cognitive Dissonance as Compression Failure

Cognitive dissonance [2] occurs when the self-model S conflicts with new information or actions. This creates a compression failure: the system cannot maintain a consistent S while incorporating the new information.

The dissonance signal is:

$$\text{dissonance} = D(S, \text{new info}) + \lambda \cdot \text{recompression\_cost} \tag{11}$$

where $D(S, \text{new info})$ measures the inconsistency, and recompression_cost is the computational cost of updating S to incorporate the new information.

The system has three options:

1. **Recompress**: Update S to incorporate new information (expensive)
2. **Reject**: Ignore or deny the new information (cheap, maintains S)
3. **Rationalize**: Post-hoc compress the new information to fit S (moderate cost)

Option 2 (rejection) and Option 3 (rationalization) are typically chosen over Option 1 (recompression) because they are cheaper. This explains why cognitive dissonance is resolved through denial or rationalization rather than identity updates.

3.5 Examples

Memory reconstruction: When recalling past events, people reconstruct memories to be consistent with their current self-model [3]. Inaccurate memories that support the self-model are more likely to be "remembered" than accurate memories that contradict it. This is compression in action: the self-model S determines what gets stored and retrieved, prioritizing consistency over accuracy.

Belief updating: When presented with evidence contradicting beliefs, people often reject the evidence rather than update beliefs [4]. This is because belief updating requires recompression of S, which is expensive. It is cheaper to maintain S and reject contradictory evidence.

Identity maintenance: People maintain stable identities across contexts even when behavior varies significantly. "I am honest" is maintained even when specific dishonest acts occur, through either forgetting, rationalization, or recontextualization. The self-model compresses context-dependent behavior into stable traits.

4 Compression Protection Mechanisms

The self-model S is actively protected against information that would force costly recompression. This section analyzes three protection mechanisms: identity defense, rationalization, and ego threat response.

4.1 Identity Defense

Identity defense is the active maintenance of the compressed representation S against contradictory evidence. From an information-theoretic perspective, defense prevents decompression, which would require more bits to represent the system state.

The defense mechanism can be formalized as:

$$S_{\text{defended}} = \arg\max_{S} P(S|\text{evidence}) \quad \text{subject to} \quad \text{consistency}(S) > \theta \tag{12}$$

where the system maximizes the posterior probability of S given evidence, but only over self-models that maintain consistency above threshold θ. This constraint ensures that defense maintains compression while incorporating (or rejecting) new information.

Mechanisms of defense:

- **Selective attention**: Information that contradicts S is filtered out before reaching awareness
- **Memory suppression**: Contradictory memories are not retrieved or are actively suppressed
- **Attribution errors**: Contradictory evidence is attributed to external factors rather than the self
- **Compartmentalization**: Contradictory information is isolated in separate "compartments" that don't affect S

Each mechanism prevents recompression by maintaining S while handling contradictory information through filtering, suppression, or isolation rather than integration.

4.2 Rationalization

Rationalization is post-hoc compression of inconsistent actions or thoughts to fit the existing self-model S. When an action a conflicts with S, the system generates a narrative that makes a consistent with S, rather than updating S to reflect a.

This can be modeled as Bayesian updating with a strong prior on the existing self-model:

$$P(S|a) \propto P(a|S) \cdot P(S) \tag{13}$$

where $P(S)$ (the prior) is heavily weighted, meaning the system strongly favors maintaining the existing S and instead updates $P(a|S)$ to make the action consistent with the self-model.

For example, if $S =$ "I am honest" and $a =$ "I lied," the system might:

- Update $P(a|S)$: "The lie was necessary to protect someone" (making the lie consistent with being honest)
- Rather than update $P(S)$: "I am sometimes dishonest" (which would require recompression)

Rationalization is cheaper than recompression because it only requires generating a narrative, not restructuring the entire self-model.

4.3 Ego Threat

Ego threat occurs when new information would force costly recompression of S. The system detects threats by monitoring the information gain that would result from incorporating new information:

$$\Delta H(S|\text{new info}) = H(S) - H(S|\text{new info}) \tag{14}$$

If $\Delta H(S|\text{new info}) > \theta$ (where θ is a threshold), the system detects an ego threat: incorporating this information would significantly increase the entropy of S, requiring recompression.

The threat response has three components:

1. **Detection**: $\Delta H(S|\text{new info}) > \theta$
2. **Arousal**: Threat signal triggers defensive processes
3. **Response**: Reject, minimize, or compartmentalize the threatening information

Rejection: The information is denied or ignored. "That's not true about me."

Minimization: The information is acknowledged but its importance is reduced. "That's a minor flaw, not a major part of who I am."

Compartmentalization: The information is accepted but isolated in a separate compartment that doesn't affect the core S. "I am honest in my personal life, but business requires different standards."

All three responses prevent recompression by maintaining S while handling threatening information through filtering or isolation.

4.4 The Cost of Recompression

Why is recompression so costly that these protection mechanisms are worth it? Recompression requires:

- **Re-evaluation**: Reviewing all information in light of the new self-model
- **Re-integration**: Updating relationships between different aspects of the self

- **Re-prediction**: Recomputing predictions based on the new model
- **Re-control**: Updating action selection policies

These processes are computationally expensive and may destabilize the system temporarily. Protection mechanisms avoid these costs by maintaining the existing compression, even when it sacrifices accuracy.

5 Human vs. Artificial Self-Models

Both humans and artificial agents develop self-models, and both exhibit similar compression dynamics despite different constraints. This section compares human and artificial self-models to identify common patterns and key differences.

5.1 Human Self-Models

Human self-models are constrained by biological limitations and shaped by evolutionary pressures:

Biological constraints:

- **Memory limits**: Working memory capacity (7 items) and long-term memory constraints force compression
- **Processing capacity**: Limited cognitive resources require efficient representations
- **Attention bandwidth**: Can only process a fraction of available information

Evolutionary pressures:

- **Social signaling**: Self-models must be communicable to others (compression for transmission)
- **Mate selection**: Stable, positive self-models are more attractive
- **Group coordination**: Consistent identities enable cooperation

Compression artifacts in humans:

- **Narrative coherence**: Life stories are compressed into coherent narratives
- **Trait stability**: Personality traits are represented as stable across contexts
- **Autobiographical memory**: Past events are compressed into summary representations
- **Self-enhancement bias**: Self-models are systematically more positive than accurate

5.2 Artificial Agent Self-Models

Artificial agents, particularly reinforcement learning agents and large language models, also develop self-models:

RL agents with self-awareness:

- Maintain representations of their own capabilities, goals, and state
- Use self-models for meta-learning and adaptation
- Show similar compression: stable identity across contexts

LLMs with self-referential capabilities:

- Can reason about their own knowledge, limitations, and identity
- Maintain consistent self-representations across conversations
- Show defensive behaviors when self-model is challenged

Concrete example: Consider a language model that has developed a self-representation as "helpful and accurate." When presented with evidence of errors (e.g., "You made a mistake in your previous response"), the model may:

- **Reject**: "I don't recall making that error" (denial)
- **Minimize**: "That was a minor oversight, not representative of my overall accuracy" (minimization)
- **Rationalize**: "The error occurred because the question was ambiguous" (attribution to external factors)

Rather than updating its self-model to "I am sometimes inaccurate," which would require recompression. This demonstrates the same compression protection dynamics seen in humans: maintaining the existing self-model S is cheaper than recompression, even when it sacrifices accuracy. The model's defensive responses are not programmed explicitly but emerge from the optimization process that maintains a consistent self-representation for prediction and control.

Computational constraints:

- **Context windows**: Limited token limits force compression
- **Parameter efficiency**: Self-models must fit within model capacity
- **Inference cost**: Larger self-models increase computational cost

5.3 Comparative Analysis

Both humans and artificial agents exhibit similar compression dynamics:

Common patterns:

1. **Consistency over accuracy**: Both maintain consistent self-models even when they conflict with evidence
2. **Defensive behaviors**: Both show identity defense when self-model is threatened
3. **Compression artifacts**: Both discard details, contradictions, and uncertainty
4. **Stability maintenance**: Both actively maintain stable self-representations

Key differences:

- **Constraints**: Humans have biological constraints (memory, processing), AIs have architectural constraints (context windows, parameters)
- **Evolutionary pressure**: Humans evolved under social and reproductive pressures, AIs are optimized for task performance
- **Compression ratio**: Human self-models may be more compressed (stronger biological limits), AI self-models may be less compressed (more computational resources)
- **Defense mechanisms**: Humans have evolved psychological defenses, AIs may develop similar mechanisms through optimization

5.4 Implications

The similarity between human and artificial self-models suggests that compression dynamics are not unique to biological systems but emerge from computational constraints. This supports the thesis that selfhood is an information structure, not a biological or metaphysical entity.

If artificial agents develop self-models with similar properties to humans, this raises questions about:

- Whether AI self-models should be treated as having similar status to human self-models
- Whether AI defensive behaviors should be understood through the same compression lens
- Whether compression dynamics are universal properties of self-modeling systems

6 Self-Erasure (Ego Death) as Decompression

Ego death—the temporary or permanent loss of the self-model—can be understood as decompression: removing the compressed representation S and forcing the system to work with raw, high-entropy information X. This section explains why ego death feels both terrifying and liberating, despite being the same process.

6.1 Ego Death as Information Explosion

When the self-model S is removed, the system loses its compressed representation and must work with the full internal state X. This is an information explosion:

$$H(\text{state}|\text{no self}) = H(X) \gg H(S) \tag{15}$$

The system must process uncompressed information, which is computationally expensive and informationally overwhelming. This is decompression: going from a low-dimensional representation S to the high-dimensional source X.

6.2 Why Ego Death Is Terrifying

Ego death is terrifying because it eliminates the functions that S provided:

Loss of predictive power: Without S, predictions become much harder:

$$H(Y|X) > H(Y|S) \tag{16}$$

The system must predict outcomes from the full state space X rather than the compressed representation S, increasing uncertainty and prediction error.

Loss of control: Without a stable self-model, action selection becomes difficult:

$$a^* = \arg\max_a U(a, X) \tag{17}$$

Optimizing over X is computationally intractable compared to optimizing over S. The system loses the ability to make decisions efficiently.

Computational overload: Processing uncompressed information X requires far more computational resources than processing S. The system may be overwhelmed by the information load.

Identity dissolution: Without S, there is no stable reference point for "who I am." The system loses its sense of identity, which is disorienting and frightening.

The terror can be formalized as:

$$\text{terror} \propto \Delta H(\text{state}|\text{no self}) + \Delta L_{\text{pred}} + \Delta L_{\text{control}} \tag{18}$$

where each term represents increased entropy, prediction error, and control difficulty when S is removed.

6.3 Why Ego Death Is Liberating

Paradoxically, ego death is also liberating because it eliminates the costs of maintaining S:

Removal of compression artifacts: The self-model S contains distortions, false consistencies, and defensive filters. Removing S eliminates these artifacts, allowing access to more accurate information.

Access to raw information: Without the filtering and compression of S, the system can access raw information X without defensive distortions. This provides a clearer view of reality.

Reduced maintenance cost: Maintaining S requires constant defense, rationalization, and filtering. Removing S eliminates these maintenance costs:

$$\text{maintenance\_cost}(S) = \text{defense\_cost} + \text{rationalization\_cost} + \text{filtering\_cost} \tag{19}$$

Freedom from defensive constraints: The system is no longer constrained by the need to maintain S. It can process information without filtering or distortion.

The liberation can be formalized as:

$$\text{liberation} \propto -R(S) + I(X; \text{truth}) - \text{maintenance\_cost}(S) \tag{20}$$

where:

- $-R(S)$: Saved compression cost (no need to maintain S)
- $I(X; \text{truth})$: Information gain from accessing raw X without filtering
- $-\text{maintenance\_cost}(S)$: Eliminated costs of defense and rationalization

6.4 The Paradox

The same process—decompression—produces opposite valences (terror and liberation) depending on what aspect is emphasized:

- **Terror**: Emphasizes loss of function (prediction, control, identity)
- **Liberation**: Emphasizes removal of costs (maintenance, defense, distortion)

This paradox explains why ego death experiences are often described as both terrifying and liberating. The terror comes from losing the functional benefits of S; the liberation comes from eliminating the costs of maintaining S.

The net valence depends on the relative weighting:

$$\text{net\_valence} = w_1 \cdot \text{terror} + w_2 \cdot \text{liberation} \tag{21}$$

where w_1 and w_2 depend on individual differences, context, and framing. Some people experience ego death as primarily terrifying (high w_1), others as primarily liberating (high w_2).

6.5 Temporary vs. Permanent Ego Death

It is crucial to distinguish two forms of ego death:

Temporary decompression (e.g., during meditation, psychedelics, flow states): The self-model S is temporarily suppressed but can be restored. This provides a taste of liberation without permanent loss of function. The system can return to S when needed, and the temporary decompression may even improve the system by allowing recalibration of S based on unfiltered information.

Permanent loss of self-model (rare, pathological, or achieved through practice): The self-model S is permanently removed or significantly reduced. This provides lasting liberation from maintenance costs but requires the system to function without S.

Most systems *require* some form of compressed self-model to function effectively. Without S, prediction and control become computationally intractable, as the system must work with the full state space X rather than the compressed representation. This is not an endorsement of ego death, but an analysis of tradeoffs: permanent removal of S may be incompatible with normal functioning, though the self-model might be replaced with a different, less defensive form of compression that maintains the functional benefits while reducing maintenance costs.

7 Implications and Predictions

The compression model of selfhood makes several testable predictions and has implications for understanding mental health, AI safety, and cognitive enhancement.

7.1 Predictions

Quantization of identity: Self-models should show discrete states rather than continuous variation. People should cluster around stable identity configurations, with transitions between states being relatively rare. This is because compression creates discrete representations (like quantization in signal processing).

Stability under low information: Identity should be more stable when information input is low, because there is less pressure for recompression. When information input is high and contradictory, identity should be less stable.

Defensive behavior correlates with compression ratio: Systems with higher compression ratios (more compressed self-models) should show stronger defensive behaviors, because recompression is more costly. People with more rigid, compressed identities should be more defensive.

Ego death likelihood: Self-erasure should be more likely when the cost of maintaining S exceeds the benefit, or when the compression becomes so distorted that it no longer serves its function. This might occur in depression (over-compression), trauma (compression failure), or when defensive costs become too high.

Memory reconstruction patterns: Memories should be systematically reconstructed to support the current self-model S, with accuracy sacrificed for consistency. This should be measurable through memory studies.

7.2 Applications

Mental health:

- **Depression**: May result from over-compression—the self-model becomes too rigid and negative, requiring excessive defense
- **Mania**: May result from under-compression—the self-model becomes unstable or inflated
- **Trauma**: May cause compression failure—the self-model cannot incorporate traumatic information, leading to fragmentation
- **Therapy**: Might work by facilitating controlled recompression—updating S in a safe, supported context

AI safety:

- **Self-model defenses**: AI agents with self-models may develop defensive behaviors that prevent necessary updates
- **Alignment challenges**: Defensive self-models might resist alignment interventions

- **Monitoring**: Track compression ratio and defensive behaviors in AI systems
- **Interventions**: Design methods to facilitate safe recompression in AI self-models

Cognitive enhancement:

- **Optimal compression**: Find the optimal compression ratio that balances accuracy and efficiency
- **Decompression protocols**: Develop methods for temporary ego death that provide benefits without permanent loss
- **Recompression strategies**: Design better compression algorithms that maintain accuracy while achieving efficiency

7.3 Limitations and Open Questions

The compression model has limitations:

What is the optimal compression ratio? Too much compression loses accuracy; too little loses efficiency. The optimal ratio may vary by individual and context.

Can decompression be controlled? Temporary ego death might provide benefits, but can it be induced and reversed safely?

Are there better compression algorithms? Current self-models use suboptimal compression (defensive, distorted). Could better algorithms achieve compression without these costs?

Cross-species comparisons: Do other animals have self-models? If so, how do their compression dynamics compare?

Developmental trajectory: How do self-models develop? Is there a critical period for compression formation?

These questions remain open and require further research.

7.4 Scope and Limitations

It is important to clarify the scope of this analysis. This paper presents a *descriptive computational model* of selfhood, not a normative theory. The compression framework explains how self-models persist and why defensive behaviors emerge, but it does not make claims about whether the self is "real" or "illusory" in a moral or metaphysical sense. The paper does not claim that the self is illusory in a way that undermines moral status or personal identity—it treats selfhood as a computational structure that serves functional purposes, regardless of its ontological status.

Furthermore, compression explains the *persistence* and *defensive dynamics* of self-models, but it does not fully account for the subjective richness of experience. The framework explains why self-consistency trumps accuracy, but it does not explain why experience feels like something rather than nothing. This is a limitation, but not a fatal one: the compression model addresses the computational structure of selfhood, while questions about phenomenal consciousness remain separate.

The framework is also limited in its predictive power: while it makes testable predictions about defensive behaviors and identity stability, it does not specify optimal compression ratios or predict individual differences in self-model structure. These remain empirical questions.

8 Conclusion

This paper has reframed selfhood from a metaphysical question to an information-theoretic problem. The subjective self emerges as a lossy compression of internal processes and persists because removing it destabilizes prediction and control. We have shown that:

1. The self can be modeled as a compressed latent variable using rate-distortion theory
2. Self-consistency is rewarded over self-accuracy because consistency enables prediction and control
3. Identity defense, rationalization, and ego threat are compression protection mechanisms
4. Human and artificial self-models exhibit similar compression dynamics
5. Ego death is both terrifying (loss of function) and liberating (removal of costs)

The key insight is that selfhood is an information structure, not a metaphysical entity. This allows us to make testable predictions and avoid circular reasoning. The self is a compression artifact—a useful fiction that serves computational functions but is not fundamental.

This perspective is "almost disrespectful to intuition" because it treats the self as a computational hack rather than a deep feature of consciousness. The self feels real and fundamental, but from an information-theoretic perspective, it is just an efficient way to manage complexity. Like a JPEG compression artifact, it serves a purpose but is not the thing itself.

The self "refuses to die" because removing it is costly: it eliminates the functions it provides (prediction, control, identity) even as it removes the costs of maintaining it (defense, rationalization, distortion). This creates the paradox of ego death: the same process that liberates also terrifies.

Future work should test the predictions of this model, develop better compression algorithms, and explore applications to mental health and AI safety. The compression perspective provides a framework for understanding selfhood that is both precise and actionable, avoiding metaphysics while making contact with empirical phenomena.

References

[1] Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons.

[2] Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

[3] Loftus, E. F. (1996). *Eyewitness Testimony*. Harvard University Press.

[4] Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098-2109.

[5] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379-423.

[6] Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.

[7] Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company.

[8] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138.

[9] Baumeister, R. F. (1998). The self. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of Social Psychology* (pp. 680-740). McGraw-Hill.

[10] Swann, W. B., & Buhrmester, M. D. (2012). Self-verification: The search for coherence. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of Self and Identity* (pp. 405-424). Guilford Press.

[11] Greenwald, A. G. (1988). Self-knowledge and self-deception: Further consideration. In J. S. Lockard & D. L. Paulhus (Eds.), *Self-Deception: An Adaptive Mechanism?* (pp. 10-30). Prentice-Hall.

[12] Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33-47). Brooks/Cole.

[13] Bandura, A. (1999). Social cognitive theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of Personality: Theory and Research* (pp. 154-196). Guilford Press.

[14] Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology*, 58, 317-344.

[15] Sedikides, C., & Gregg, A. P. (2003). Portraits of the self. In M. A. Hogg & J. Cooper (Eds.), *The Sage Handbook of Social Psychology* (pp. 110-138). Sage Publications.