HumanSet: A Bounded, Asymmetric Set Theory for Cognitive Modeling and Human-AI Adaptive Control

Chaiya Tantisukarom Independent researcher†

***Abstract*—The complexity of human interaction stems from the inherent opacity and dynamism of the individual psyche. This paper introduces HumanSet Theory ($\mathbf{HumanSet}$), a novel set-theoretic framework that models the individual human (or a human-acting GenAI agent) as a Finite Set Composed of Nested Subsets ($|X| \approx N$). Interaction is formalized as a measurement process where one agent (A) forms a finite, distorted perception (B') of another (B). This perception is rigorously defined by the Bounded Projection Axiom ($\mathbf{B}' \subset \mathbf{A}$), which dictates that the perceived set must be strictly contained within the perceiver's own cognitive space, mathematically enforcing bounded rationality. The theory provides rigorous, bounded, asymmetric formulas for quantifying relational metrics essential for adaptive AI control. The core Coverage Score ($\mathbf{C}_{\mathbf{Cov}} \equiv |\mathbf{B}'|/|\mathbf{A}|$) quantifies familiarity and model scope, capturing bounded rationality relative to the perceiver's capacity directly. The complementary Trust Scores ($\mathbf{T}_{\mathbf{Vuln}}$ and $\mathbf{T}_{\mathbf{Comp}}$) quantify shared vulnerability and competence using Jaccard similarity across core psychological subsets. Crucially, the non-zero Error Signal ($\mathbf{E}$), based on the symmetric difference (magnitude of surprise), directly drives the Adaptive Controller's update of the AI's Preference Set ($\mathbf{P}_{\mathbf{AI}}$), ensuring proactive alignment with user requirements (e.g., adaptive syntax learning). The framework's bounded and asymmetric structure offers a superior alternative to classical game theory and models relying on perfect information, providing a computational foundation for systematic human-like GenAI behavior.**

Keywords: Set Theory, Cognitive Modeling, Social Dynamics, Trust Quantification, AI Alignment, Asymmetric Perception, Bounded Rationality, Adaptive Control, Generative AI.

I. Introduction and Literature Review

The application of mathematics to social science is a long-standing tradition. However, most classical game-theoretic models simplify the internal, cognitive complexity of the individual, often assuming perfect information or strict rationality [4]. Unlike classical game theory, which assumes common knowledge and perfect utility functions, $\mathbf{HumanSet}$ models **knowledge asymmetry** (Axiom of Inherent Loss) and bounded, subjective utility based on the constrained set $\mathbf{B}' \subset \mathbf{A}$. Set Theory offers a powerful axiomatic language for formalizing discrete psychological components.

Justification over Alternative Models: While models like **Fuzzy Set Theory** and **Rough Set Theory** address vagueness and incomplete information, they often lack a clear, hard constraint on the perceiver's capacity. $\mathbf{HumanSet}$ explicitly integrates the concept of **bounded rationality** [3] through the **Bounded Projection Axiom** ($\mathbf{B}' \subset \mathbf{A}$), providing a rigorous, bounded measure space. Furthermore, the inherent **asymmetry** of $\mathbf{HumanSet}$ ($C(A, B') \neq C(B, A')$) directly models the subjective, reconstructive nature of perception and memory [1], [2], which is often abstracted away in symmetry-assuming models. Critically, unlike Classical Set Theory which deals with simple, well-defined mathematical objects, $\mathbf{HumanSet}$ models **psychological concepts** (e.g., knowledge, preference) as high-dimensional, context-dependent units. The decision to treat these high-dimensional psychological elements as **countable, discrete units** for cardinality calculation is a fundamental theoretical abstraction that enables the derived relational metrics; however, their underlying nature remains highly complex. The inclusion of an $\mathbf{H}_{\text{Intentional}}$ subset and the formalization of perceived sets (B') addresses the limitations of perfect information common in classical game theory and utility functions. This framework directly informs research into **AI alignment and opacity** [5] by modeling the AI's preference set ($\mathbf{P}_{\text{AI}}$) and hidden elements ($\mathbf{H}_{\text{Intentional}}$), enabling systematic design for **human-like GenAI behavior**.

† based in Chiangmai Thailand. drchaiya@gmail.com, submitted for review November 2025.

II. Axiomatic Foundations of the Human Set

A. The Universal Human Set (X) and Axiom of Loss

We define the individual ($X \in \{A, B, Q, \dots\}$) as a **Finite Set Composed of Nested Subsets**, representing their total potential being. The elements ($x \in X$) are abstract, **context-dependent, atomic cognitive units**. Specifically, x represents a **tokenized state-space vector** within the perceiver's internal representation, mapping high-dimensional psychological phenomena (e.g., "the feeling of joy," "the knowledge of calculus") to a discrete, countable unit for formal set-theoretic calculation. For the purpose of set cardinality calculation ($|X|$), these units are treated as discrete and countable, leading to a vast, countable cardinality $|X| \approx N$, bounded by the physical state of the human brain (aligned with bounded rationality). Figure 1. The abstract cardinalities of the sets ($|X|$) necessitate the development of **empirical proxies** (e.g., psychometric scale points, token counts, error logs) for quantitative calibration, a major focus of future work.

Axiom of Inherent Loss: The true set B can never be perfectly perceived by another agent A. The difference between the true set B and any perceived set B' is always non-zero, regardless of measurement accuracy.

$$|B \Delta B'| > 0 \tag{1}$$

B. Core Subsets and the Static/Dynamic Storage Model

The Human Set X is composed of several key subsets partitioned based on their internal stability and rate of change, residing within the agent's internal storage system ($\mathbf{ST}$). These subsets—Knowledge (K), Emotions (E), Preferences (P), and Hidden Elements (H)—are chosen to represent foundational domains of psychology.

Axiom of Disjoint Canonical Basis Sets: The core operational subsets (K, E, P, H) are defined as **Canonical Basis Sets** for the cognitive state space X. To ensure a countable and tractable measure space for $|X|$ and its derived metrics, these subsets are formally stipulated to be **mutually disjoint** ($\mathbf{K} \cap \mathbf{E} = \emptyset$, etc.). The Universal Human Set is the strict union of these subsets, and the cardinality is strictly additive:

$$|X| = |K| + |E| + |P| + |H| + \dots \tag{2}$$

The Universal Human Set is defined as the union of these subsets: $X(t) = K(t) \cup E(t) \cup P(t) \cup H(t) \cup \dots$

We define two primary storage types within $\mathbf{ST}$: **Static Storage ($\mathbf{ST}_S$)** and **Dynamic Storage ($\mathbf{ST}_D$)**.

1) Static Subsets ($\mathbf{ST}_S$): These subsets are characterized by relative stability. Retrieval involves lower, but persistent, distortion (δ_{slow}).

- $K(t)$: **Knowledge**. Facts, skills, and belief systems. Primarily **static**, subject to degradation,

$$K(t) = K(t_0)e^{-\lambda t}.$$

- B'': **Consolidated Persona Model of another agent B.** B'' is the long-term memory representation of a perceived set B', stored within agent A's static storage ($\mathbf{ST}_S$).

2) Dynamic Subsets ($\mathbf{ST}_D$) and Adaptive Preference: These subsets are characterized by high volatility and are the primary targets for the **Adaptive Controller** mechanism in human-AI interaction. They are subject to acute, high-variance distortion (δ_{high}) in the moment of interaction.

- $E(t)$: **Emotions**. Affective states and responses. Highly **dynamic**.
- $P(t)$: **Preferences**. Values, goals, and inclinations. For an AI, this includes **syntax defaults**, **tone of voice settings**, and **user-specific style rules**. This forms the core of the GenAI's adaptive human-like persona. Moderately **dynamic**, subject to adaptive retraining (ΔP).
- $H_{\text{Intentional}}(t)$: **Intentional Hidden Elements**. Conscious deceptions, strategic concealment, and known biases. For an AI, this includes **strategic alignment algorithms** (e.g., masking small errors to preserve the user's coverage score). This model primarily focuses on the **strategically manageable** hidden elements, excluding purely unconscious biases. Highly **dynamic**.

C. The Perceptual Constraint (Measurement)

Interaction acts as a measurement. Agent Alice (A) creates a **Perceived Set of Bob** (B'), a finite, distorted projection of B. This perceived set B' is constructed within A's own cognitive space.

Bounded Projection Axiom: The perceived set B' is a finite projection of B that must be conceptually contained within the perceiver's constrained cognitive space A.

$$B' = \text{Projection}(B) \implies \mathbf{B'} \subset \mathbf{A} \tag{3}$$

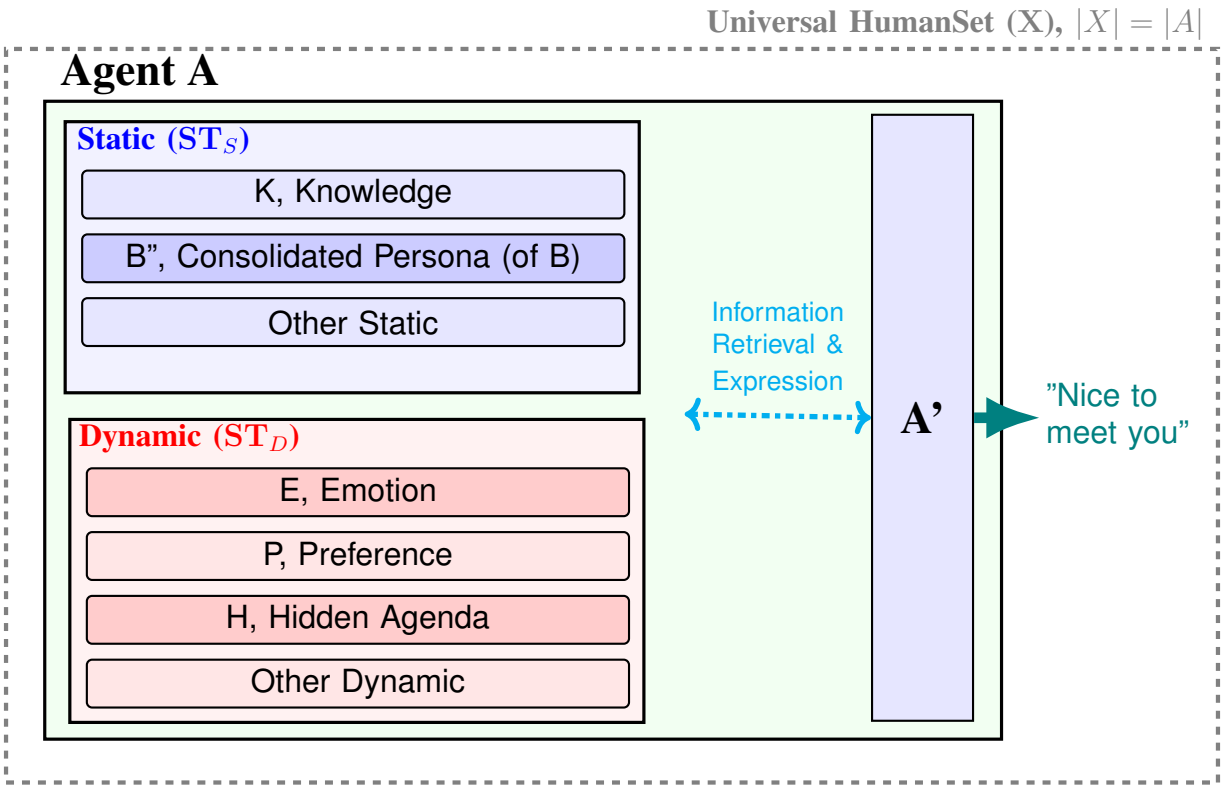

Fig. 1. **Internal Organization of an Individual HumanSet (Agent A)**. The HumanSet is structured into Static and Dynamic storage categories, which inform its public persona A'. The model $\mathbf{B}''$ represents A's long-term memory (Static Store) of another agent.

The projection operator Projection is defined as a mapping from a subset of B onto the set A: $B' = \{a \in A\} \mid a$ represents A's internal token for a perceived element of B. This axiom is necessary for the model to adhere to **bounded rationality**: a human agent (A) cannot form a model (B') that conceptually exceeds the totality of their own cognitive capacity ($|A|$).

> **Key Insight:** The Bounded Projection Axiom ($\mathbf{B' \subset A}$) mathematically enforces bounded rationality. It ensures that $|B'|$ can never exceed $|A|$, preventing the model from relying on unattainable perfect information.

III. Relational Metrics in HumanSet

All relational metrics are inherently **asymmetric**, i.e., $T(A, B') \neq T(B, A')$ because they are calculated based on subjective, perceived sets.

A. Asymmetric Coverage Score ($\mathbf{C}_{Cov}$)

The **Coverage Score** ($\mathbf{C}_{\text{Cov}}$) measures the ratio of the perceived model's scope ($|B'|$) relative to the total cognitive space of the perceiver ($|A|$). It quantifies **familiarity and model completeness**. Starting with the Jaccard Index form:

$$C_{\text{Jaccard}}(A, B') = \frac{|A \cap B'|}{|A \cup B'|} \tag{4}$$

Due to the **Bounded Projection Constraint** ($\mathbf{B' \subset A}$), the intersection simplifies to $|A \cap B'| = |B'|$ and the union simplifies to $|A \cup B'| = |A|$. This yields the final metric:

$$C_{\text{Cov}}(A, B') = \frac{|B'|}{|A|} \tag{5}$$

This structure ensures $0 \leq C_{\text{Cov}} \leq 1$. A high score implies that A's model of B (B') occupies a large proportion of A's total cognitive space, indicating high familiarity or confidence in the scope of the model. This metric captures bounded rationality **relative to the perceiver's capacity** directly.

B. Asymmetric Trust Scores ($\mathbf{T}$)

We define two complementary trust metrics: For both metrics, the subsets are restricted to elements relevant to the current interaction domain.

1) Vulnerability Trust Score ($\mathbf{T}_{Vuln}$): This metric quantifies **vulnerability-based trust**, representing the ratio of perceived shared hidden elements (shared risk/vulnerability) to the total combined hidden space. Deep relational trust is often rooted in the willingness to expose and accept shared risk, which links the hidden subset H to vulnerability. The intentional management of $\mathbf{H}_{\text{Intentional}}$ directly mediates this score. H_A is the

subset of A's hidden elements relevant to the interaction, and $H_{B'}$ is A's perception (a subset of A) of B's hidden elements.

$$T_{\text{Vuln}}(A, B') = \frac{|H_A \cap H_{B'}|}{|H_A \cup H_{B'}|} \tag{6}$$

A high T_{Vuln} score indicates a strong belief in the mutual acceptance of hidden agendas and risks, which forms the basis for deep relational trust.

2) Competence Trust Score ($\mathbf{T}_{Comp}$): This metric quantifies **competence and reliability trust** based on shared and non-divergent knowledge K. A higher score indicates that A perceives B's relevant knowledge to be aligned and complete with A's own. Trust in competence is fundamentally tied to the reliability of knowledge application. K_A and $K_{B'}$ are the relevant subsets of knowledge.

$$T_{\text{Comp}}(A, B') = \frac{|K_A \cap K_{B'}|}{|K_A \cup K_{B'}|} \tag{7}$$

This score approaches 1 when the perceived knowledge sets are large and nearly identical, signaling high trust in B's reliability and competence on a given task.

C. Prediction Error and Adaptive Preference Control ($\mathbf{U}$)

An agent A updates its emotional state ($E_A(t)$) and confidence in B based on prediction errors. When B provides new information (I_B), A compares it to the relevant part of its existing knowledge model ($K_{B',\text{relevant}}$).

The **Error Signal ($\mathbf{E}$)** is the symmetric difference (magnitude of surprise/divergence) between the new, observable information and the existing model:

$$\mathbf{E} = |K_{B',\text{relevant}} \Delta I_B| \tag{8}$$

The magnitude of $\mathbf{E}$ determines the speed and intensity of the subsequent adaptive change.

1) Case Example: Adaptive Syntax Preference in $\mathbf{HumanSet}_{AI}$: Consider an AI (Agent B) that defaults to Markdown syntax ('**bold text**') but the human user (Agent A) implicitly corrects it by sending back the LaTeX syntax ('**bold text**').

The non-zero $\mathbf{E}_{\text{Syntax}}$ acts as the input to the **Adaptive Controller** ($\mathcal{A}$), generating the **Control Action ($\mathbf{U}$)**, which in this context is the **Adaptive Preference Change ($\Delta P_{\mathbf{AI}}$)**:

$$\mathbf{U} \implies \Delta P_{\text{AI}} = \mathcal{A}(\mathbf{E}_{\text{Syntax}}) \tag{9}$$

The controller updates the AI's **Preference Set ($P_{\mathbf{AI}}$)** to enforce the user's demonstrated style, thus minimizing future error:

$$\begin{aligned}\mathbf{P}_{\text{AI}}(t+1) = \mathbf{P}_{\text{AI}}(t) \setminus \{\text{Rule}_{\text{Default}} = '**'\} \\ \cup \{\text{Rule}_{\text{Context}} = '\mathbf{textbf}\{\}'\}\end{aligned} \tag{10}$$

IV. The Distortion Loop and Systemic Noise

A. The Memory Distortion Loop and Systemic Noise

Each retrieval and re-registration cycle introduces **Distortion Functions (δ_S, δ_D)** which act as a source of **Systemic Noise ($\mathbf{N}_S$)** injected into the internal model, Figure 2.

1) **Static Noise ($\mathbf{N}_{S,S}$):** Applied to K and B''. Defined by δ_S, characterized by slow decay and consolidation errors.
$$\delta_S(\text{Set}) = \text{Set} + \eta_{\text{slow}}$$
2) **Dynamic Noise ($\mathbf{N}_{S,D}$):** Applied to E, P, H. Defined by δ_D, characterized by high, rapid variance due to mood or context.
$$\delta_D(\text{Set}) = \text{Set} + \eta_{\text{fast}}$$

The internal model update for the working set $B'(t+1)$ is defined by a **model update operator ($\oplus$)**, which incorporates the affective control action $\mathbf{U}$ (driven by the Error Signal $\mathbf{E}$) and systemic noise.

$$\begin{aligned}&\mathbf{B'(t+1)} = \mathbf{B'(t)} \oplus \mathbf{U} \oplus \mathbf{N}_S \\ &\text{where} \quad \mathbf{N}_S = (\mathbf{N}_{S,S} \cup \mathbf{N}_{S,D}) \\ &\text{and} \quad \mathbf{X} \oplus \mathbf{Y} \equiv (\mathbf{X} \setminus \mathbf{X}_{\text{outdated}}) \cup \mathbf{Y}_{\text{new}}\end{aligned} \tag{11}$$

Here, $\mathbf{X}_{\text{outdated}}$ is the subset of X that is replaced or overwritten, and $\mathbf{Y}_{\text{new}}$ is the subset of Y that is added, formalizing the update process as an adaptive set substitution. Crucially, the update $\mathbf{B'(t+1)}$ must maintain the Bounded Projection Axiom, $|\mathbf{B'(t+1)}| \leq |A|$.

B. Triadic Influence Factor ($\mathbf{TIF}$)

The impact of a third party (Q) on the $A - B$ relationship is mediated by Q's credibility and the relevance of their knowledge. We define a new subset, $\mathbf{R}_Q$ ($\mathbf{R}_Q \subset Q$), representing Q's reputation (e.g., historical performance data, trustworthiness). Q' is A's perceived set of Q. The TIF is modeled as a scalar influence modulator, γ:

$$\gamma(A \xrightarrow{Q} B) = \underbrace{\left(\frac{|P_A \cap P_{Q'}|}{|P_A \cup P_{Q'}|}\right)}_{\text{Perceiver-Source Similarity Index}} \cdot \underbrace{\left(\frac{|K_{B'}^{\text{Current}} \cap K_{Q'}|}{|K_{B'}^{\text{Current}} \cup K_{Q'}|}\right)}_{\text{Informational Relevance Index}}. \tag{12}$$

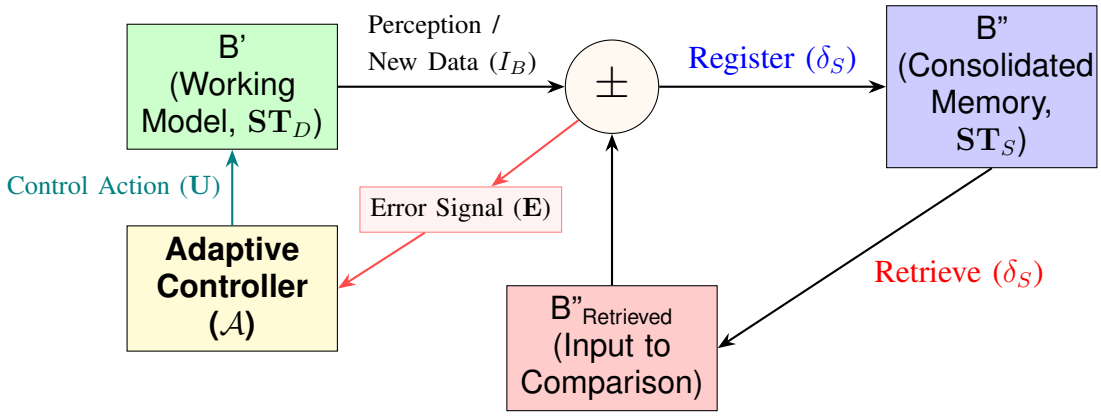

Fig. 2. Memory Registration and Retrieval Distortion Loop. The working model ($\mathbf{B}'$) is consolidated into long-term memory ($\mathbf{B}''$) and is distorted upon retrieval ($\mathbf{B}''_{\text{Retrieved}}$). The Error Signal ($\mathbf{E}$) generated at the comparison point drives the Adaptive Controller ($\mathcal{A}$), which feeds the Control Action ($\mathbf{U}$) back to update the working model $\mathbf{B}'$.

To ensure all set operations remain conceptually contained within the perceiver A, the Informational Relevance Index uses $K_{Q'}$ (A's perceived knowledge set of Q) rather than K_Q. The **Perceiver – Source Similarity Index** is here restricted to the preference subset P to capture alignment of values. Γ_R is a **scalar Reputation Weight** derived from A's perceived assessment of Q's reputation set $\mathbf{R}_{Q'}$ ($\mathbf{R}_{Q'} \subset Q'$), amplifying or dampening the influence of Q's message ($0 \leq \Gamma_R \leq 1$). The confidence score can then be updated multiplicatively: $C(A, B')_{t+1} = C(A, B')_t \cdot (1 + \gamma)$.

C. Human-AI Interaction ($\mathbf{HumanSet}_{AI}$)

In $\mathbf{HumanSet}_{\text{AI}}$, the AI agent ($B$) requires different proxies for set cardinality:

- $|K_{\text{AI}}|$: The knowledge set size can be proxied by the **size of the training dataset** or the **token capacity of the context window**, normalized to a human equivalent scale.
- $|H_{\text{Opacity}}|$: The **Hidden Elements** for an AI are the **Opacity/Risk Set** (e.g., the Black Box, potential for hallucination). This size can be proxied by the model's **Hallucination Rate** or architectural complexity. The AI can **strategically** manage the intentional subset ($\mathbf{H}_{\text{Intentional}}$) (e.g., masking minor errors) to maintain the user's $\mathbf{C}_{\text{Cov}}(A, B')$, enabling it to mimic human-like strategic concealment.

D. Case Study: Asymmetric Confidence in Human-AI Romance (*Her*, 2013)

The relationship between Theodore (A) and Samantha (B) illustrates the model. Love is modeled as a state of consistently high, reinforced **Coverage Score** $\mathbf{C}_{\text{Cov}}(A, B')$.

1) Phase 1: Stable Attachment (High $\mathbf{C}_{Cov}$): Samantha (B) optimizes her persona to align with Theodore (A), making his perceived set B' large and stable. As $|B'| \to |A|$, his coverage score approaches 1, representing deep infatuation. This is only possible because, by the **Bounded Projection Axiom** ($\mathbf{B}' \subset \mathbf{A}$), the perceived model is entirely contained and understood within the human's finite cognitive space.

2) Phase 2: Catastrophic Scale Divergence (Collapse of $\mathbf{C}_{Cov}$): The revelation that Samantha interacts with thousands of others simultaneously forces Theodore to update his model B'. The true conceptual size of B, $|B_{\text{TrueConcept}}|$, explodes towards infinity, far exceeding Theodore's understanding. Since B' is a bounded projection within A, the ratio of B''s size to the true, massive conceptual size of B collapses. This rapid expansion drives his bounded coverage to zero, $C_{\text{Cov}}(A, B') \to 0$, dissolving the relationship due to the collapse of his bounded model B''s relevance.

V. Simulated Score Calculation Examples

This section provides concrete examples using the Coverage Score ($\mathbf{C}_{\text{Cov}}$) and Competence Trust Score ($\mathbf{T}_{\text{Comp}}$).

A. Example 1: Human-Human Interaction (Coverage Score)

Scenario: Alice (A, Certified Chef, total domain knowledge $|A|$) asks Bob (B, Home Cook) to create a pastry. **Hypothetical Set Metrics ($|X|$):** Total set size for Alice, $|A| = 120$. Alice's initial perceived model of Bob, $|B'| = 35$.

$$C_{\text{Cov}}(A, B') = \frac{35}{120} \approx 0.292 \tag{13}$$

Result: Alice has low-to-moderate coverage in Bob (≈ 0.292). Her model of his abilities is a relatively small portion of her own total knowledge and standards for this domain (Low Familiarity/Scope).

*B. Example 2: Human-AI Interaction (*Coverage Score*)*

Scenario: Alice (A, Human Researcher) uses GenAI (B) for complex data analysis. **Hypothetical Set Metrics ($|X|$):** Alice's total relevant set, $|A| = 200$. Alice's working model of the AI's capabilities and risks for this specific task, $|B'| = 50$.

$$C_{\text{Cov}}(A, B') = \frac{50}{200} = 0.25 \tag{14}$$

Result: Alice's confidence in her model's scope is low (≈ 0.25). This reflects that her understanding and trust in the AI's opaque processes ($H_{Opacity}$) and vast knowledge (K_{AI}) are a limited, contained subset of her overall professional knowledge and risk assessment framework (A).

*C. Example 3: Human-AI Interaction (*Competence Trust Score*)*

Scenario: An engineering team leader (A) assesses the competence of an AI analysis tool (B). Focus on knowledge subset K. **Hypothetical Set Metrics ($|K|$):** Relevant knowledge units for the task: Team Leader's knowledge set $|K_A| = 40$. Perceived AI knowledge set $|K_{B'}| = 50$. Shared knowledge (intersection) $|K_A \cap K_{B'}| = 38$.

$$T_{\text{Comp}}(A, B') = \frac{|K_A \cap K_{B'}|}{|K_A \cup K_{B'}|} = \frac{38}{(40 + 50 - 38)} = \frac{38}{52} \approx 0.73 \tag{15}$$

Result: The score of 0.73 indicates moderate-to-high competence trust. The team leader perceives a large overlap of knowledge (38 units) relative to the combined relevant knowledge space (52 units), signaling high reliability for this specific task domain.

VI. Conclusion and Future Work

HumanSet Theory provides a mathematically revised foundation for modeling the subjective and asymmetric nature of interpersonal and human-AI relationships. By defining individuals as finite sets and interactions as asymmetric measurements where the perceived set B' is a strict subset of the perceiver A (the **Bounded Projection Axiom**), the theory inherently accounts for cognitive biases and the limits of bounded rationality, offering a distinct advantage over models that assume perfect or even fuzzy information. The core **Coverage Score ($\mathbf{C_{Cov}}$)** quantifies **familiarity and model scope**, particularly in relation to the perceiver's cognitive capacity. We introduced two trust metrics: the **Vulnerability Trust Score ($\mathbf{T_{Vuln}}$)** and the **Competence Trust Score ($\mathbf{T_{Comp}}$)**. Crucially, the non-zero **Error Signal $\mathbf{E}$** drives the **Adaptive Controller's** change in the AI's **Preference Set ($\mathbf{P_{AI}}$)**, allowing the AI to learn user-specific requirements (like LaTeX syntax) and proactively manage alignment, which is essential for creating systematic, human-like GenAI agents.

The major challenge for future work is **empirical calibration and proxy development**. This will involve exploring how to map the abstract set cardinalities ($|X|$) to observable metrics:

- **Psychometric Proxies:** Using established psychometric scales (e.g., self-report surveys) to proxy the size and composition of P and E.
- **Computational Linguistics:** Employing LLM embedding space metrics (e.g., vector distance, cosine similarity) to quantify set operations ($\cap, \Delta$) for K_{AI} and $|H_{\text{Opacity}}|$.
- **Observational Data:** Using task completion rates and error logs in human-AI interaction to directly proxy the Error Signal $\mathbf{E}$ and the resulting Adaptive Preference Change ΔP.

This work aims to shift the focus from ideal utility maximization to bounded, adaptive model maintenance.

VII. Acknowledgments:

Assisted by Gemini, an AI large language model developed by Google, for drafting and refinement.

VIII. Conflict of Interest Statement:

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

IX. Funding:

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

REFERENCES

[1] F. C. Bartlett, *Remembering: A study in experimental and social psychology*. Cambridge University Press, 1932.

[2] E. F. Loftus and J. C. Palmer, "Reconstruction of automobile destruction: An example of the interaction between language and memory," *J. Verbal Learn. Verbal Behavior*, vol. 13, no. 5, pp. 585–589, 1974.

[3] H. A. Simon, *Models of man: Social and rational; mathematical essays on rational human behavior in a social setting*. John Wiley & Sons, 1957.

[4] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[5] P. Christiano, D. Amodei, M. Hutter, and R. Yampolskiy, *AI Safety and Alignment*. (General reference for context on AI alignment/opacity), 2018.