What AI Is and Is Not:

Concepts, Logic, and the Limits of Human Analogy

Lloyd Axail Cordova Barrera

Independent Researcher

ORCID: 0009-0005-3424-9212

Abstract

When people describe an AI model as thinking, understanding, or hallucinating, they apply human psychological vocabulary to a mathematical system. This paper argues that doing so constitutes a category mistake with measurable consequences for how AI systems are designed, evaluated, and held accountable. The argument proceeds through four methods: conceptual analysis of the path from raw data to AI systems; critical engagement with the main philosophical positions on mind, intentionality, and machine cognition; a formal logical argument whose premises are argued rather than assumed; and annotated code that makes the computational process concrete. The paper engages the principal objections to its thesis — including Dennett's intentional stance and Hinton's observations on large-scale neural networks — and responds to each. It does not claim that AI systems are simple or that nothing philosophically significant occurs in large neural networks. It claims that the specific vocabulary of human psychology describes properties for which there is no current evidence in AI systems, and that using that vocabulary produces harms that more precise language would avoid.

Keywords: *anthropomorphism; artificial intelligence; philosophy of mind; intentional stance; category mistake; machine learning; deep learning; ethics of AI; levels of abstraction.*

1. Introduction

AI systems are described, in both technical and public discourse, using language borrowed from human psychology. A model is said to think, understand, hallucinate, or decide. These descriptions feel natural because AI systems produce outputs that resemble human behavior. But resemblance is not identity, and the vocabulary we use to describe a system shapes the expectations we bring to it, the trust we extend to it, and the accountability we assign when it fails.

This paper argues that applying human psychological predicates to current AI systems is a category mistake — the error of applying a concept to something that does not belong to the category for which that concept was defined. It is not the paper's position that AI systems are trivial or that their outputs are merely random. It is the paper's position that the specific terms thinking, understanding, intending, and hallucinating require phenomenal consciousness and genuine intentionality, and that there is no credible current evidence that any AI system instantiates those properties.

The argument is developed carefully. The main philosophical objections are engaged rather than dismissed. The paper acknowledges what is genuinely uncertain and does not claim more than the evidence supports. Its practical aim is to contribute to a more precise public and technical vocabulary for describing what AI systems actually do.

The paper is organized as follows: methodology (Section 2), literature review (Section 3), formal argument (Section 4), technical foundations (Sections 5–9), hallucinations and anthropomorphic language (Sections 10–11), ethics (Section 12), code with philosophical exegesis (Section 13), and conclusion (Section 14).

2. Methodology

This paper combines four methods. Each is chosen for a specific reason, not merely described.

The first is conceptual analysis. The core terms of the debate — intelligence, understanding, learning, hallucination — carry different meanings in their ordinary uses and their computational uses. Conceptual analysis makes those differences explicit. This method is appropriate here because many of the errors the paper addresses are errors of concept application: they arise not from false empirical beliefs but from using a term outside the domain for which it was defined.

The second is critical engagement with the philosophical literature on mind and intentionality. The question of whether machines can think or understand is not new. Substantial philosophical work has been done on it, and any argument that ignores that work is likely to repeat mistakes already identified. This paper works through the main positions — including positions that challenge its own thesis — and responds to them. The goal is not comprehensiveness but honest engagement with the strongest objections.

The third is logico-formal argumentation. The central claim is expressed as a set of premises leading to stated conclusions. This makes the logical structure explicit and separable from the rhetorical presentation, so that each step can be evaluated independently. Where premises rest on empirical claims, those claims are identified as such and distinguished from the logical derivation.

The fourth is code exegesis. Two annotated code examples — a Variational Autoencoder and a next-token prediction model — are analyzed not as programming demonstrations but as evidence. The goal is to show, at the level of actual computation, what these systems do and what they do not do. This method is appropriate because abstract philosophical claims about AI are often made without reference to how the systems actually work. The code grounds the argument in specifics.

One clarification about scope: this paper addresses current AI systems. It does not make claims about what future systems might be. If future AI systems provide credible evidence of phenomenal consciousness — through, for example, behavioral markers that currently have no non-conscious explanation, or through architectural features that current theories of consciousness identify as sufficient — the argument would need to be revisited. The paper's conclusions are bounded by the state of current systems and current evidence.

3. Literature Review

The question of whether machines can think is as old as the modern computer. Turing (1950) proposed replacing it with a behavioral test: if a machine produces responses indistinguishable from a human's, there is no operational reason to deny it intelligence. Turing's proposal is significant for this paper because it deliberately sets aside the question of inner experience. Turing did not claim machines understand — he claimed the distinction may not matter for practical purposes. This paper argues it does matter, and in ways Turing's framing did not anticipate.

3.1 Against Psychological Attribution

Weizenbaum (1976) built ELIZA, a program that simulated psychotherapeutic conversation through simple pattern-matching rules with no semantic content. He found that users — including colleagues who knew how the program worked — attributed understanding and empathy to it. Weizenbaum called this the ELIZA effect: the human tendency to project psychological states onto systems that produce human-like output. His concern was not that ELIZA was powerful but that it was not — and yet the attribution occurred anyway. Contemporary language models are vastly more sophisticated, but the mechanism Weizenbaum identified is structurally the same.

Searle (1980) argued through the Chinese Room thought experiment that syntactic manipulation — the formal operation on symbols according to rules — does not produce semantic content, which is meaning grounded in experience and intentionality. A system that processes Chinese symbols according to rules produces correct outputs without understanding Chinese. This argument has been contested. The systems reply holds that understanding belongs to the system as a whole rather than to any component. The robot reply suggests that grounding symbols in sensorimotor interaction with the world would produce genuine understanding. This paper does not treat Searle's argument as conclusive. Rather, it treats it as identifying a genuine problem: the gap between producing correct outputs and genuinely understanding the domain those outputs concern. The weight of the argument in this paper rests primarily on Nagel and Floridi rather than on Searle.

Nagel (1974) argued that consciousness has a subjective, first-person character — what philosophers call phenomenal experience — that is not captured by any third-person physical or functional description. His question — what is it like to be a bat? — points to the fact that a complete physical account of echolocation leaves out the qualitative experience of navigating by sound. Applied to AI: we can describe in complete mathematical detail what a language model does when it generates text. What we have no evidence for is that there is anything it is like to be that process. The outputs resemble the products of understanding, but resemblance at the output level does not establish similarity at the level of inner experience.

De Waal (1997, 2016) introduced the concept of anthropodenial — the mirror error of anthropomorphism — which is the denial of mental properties to entities that have them. De Waal applied this concept to animal cognition: the tendency in some scientific circles to refuse psychological attribution to animals even when the behavioral evidence supports it. This concept is relevant here as a preemptive clarification. The present paper argues against anthropomorphizing AI; it does not argue for anthropodenial. The distinction is this:

anthropodenial is an error because animals demonstrably have perceptual and affective states whose denial requires ignoring evidence. The argument against AI anthropomorphism is not symmetric: it rests on the absence of evidence for phenomenal states in AI systems, not on a blanket refusal to consider the question. If evidence of phenomenal consciousness in AI systems emerged, the position in this paper would need to change. The argument is empirically bounded, not dogmatic.

3.2 The Intentional Stance and Its Limits

Dennett (1987) argues for the intentional stance: the strategy of interpreting a system's behavior by attributing to it beliefs, desires, and rationality. We apply this stance to other people, to animals, and — usefully — to thermostats and chess programs. The claim is not that thermostats literally have desires but that attributing desires to them is an efficient explanatory strategy that accurately predicts their behavior. For Dennett, there is no deeper fact about whether the system really has those properties — the stance is valid when it works.

This is a coherent philosophical position. The present paper does not claim it is wrong as a theory of mind. It claims that the intentional stance, when applied to AI in public discourse, regulatory frameworks, and product communication, is used in contexts for which it was not designed and in which it produces specific harms. Dennett's intentional stance is a tool for explanation and prediction in a philosophical context where the question of genuine possession of mental states is bracketed. It does not provide a framework for assigning legal responsibility, calibrating user trust, or auditing model behavior — all of which require knowing not just what predictions the stance licenses but what is actually happening in the system. Using intentional language in those contexts conflates the explanatory level with the causal level in ways that obscure accountability.

3.3 Large Neural Networks: Hinton, Bender, and the Question of Scale

Two serious and partially conflicting positions on the cognitive significance of large neural networks are relevant to this paper, and it is necessary to engage both honestly rather than citing one and ignoring the other.

Hinton, Bengio, and LeCun (2015) demonstrated that deep neural networks trained on large datasets learn hierarchical representations that support remarkable generalization. Hinton has since expressed concern — in technical writing and in public statements — that large-scale neural networks may be developing internal representations that are not well described by existing frameworks, and that the field does not yet have adequate vocabulary for what is happening in these systems. This is a serious position from one of the central figures in deep learning. It deserves acknowledgment rather than dismissal.

Bender, Gebru, McMillan-Major, and Shmitchell (2021) argue the contrary: that large language models are stochastic parrots — systems that have learned the statistical structure of language without access to the grounded meaning that structure encodes in human use. Their argument is that fluency and comprehension are independent dimensions, and that scaling a system optimized for fluency does not, by itself, produce comprehension. Producing text that sounds like the text a knowledgeable person would produce is not the same as having the knowledge that would ground it.

This paper holds the following position, stated with appropriate caution: Hinton is correct that large neural networks develop internal representations that are not trivial and that current frameworks do not fully characterize. Bender et al. are correct that fluency is not comprehension and that scale does not by itself bridge the gap between statistical association and grounded meaning. The two positions are compatible if framed correctly: something interesting and not yet fully understood is happening in large neural networks, and that something is not well described by the psychological vocabulary currently applied to it. The response to Hinton's concern is not to use human psychological terms more freely but to develop more precise vocabulary for what these systems actually do. That is the position this paper defends.

3.4 Levels of Abstraction

Floridi (2004, 2008) develops the concept of levels of abstraction as a framework for resolving apparent contradictions in descriptions of complex systems. Whether a system has a given property depends on the level of abstraction at which it is observed. At the behavioral level, a language model appears to understand: it produces contextually appropriate, semantically coherent responses. At the implementation level, it multiplies matrices and samples from probability distributions. Both descriptions are simultaneously valid; they operate at different levels of abstraction. This framework is important for this paper in two ways. First, it clarifies that the paper's argument is not that behavioral descriptions of AI are false — it is that they are descriptions at the wrong level of abstraction for the purpose of ethical accountability and legal responsibility. Second, and developed further in Section 12, Floridi's framework implies that causal intervention — fixing a problem, assigning responsibility, modifying behavior — must be performed at the implementation level. Accountability requires knowing what is actually happening in the system, not just what the system appears to be doing from the outside.

3.5 The Aristotelian Frame

Aristotle's account of the soul in De Anima describes a hierarchy of capacities: nutritive (plants), sensitive (animals), rational (humans). What is philosophically important about this framework is not its specific content but its structure: it recognizes graded capacities without collapsing them into a single category and without requiring that lower capacities be described using the vocabulary appropriate to higher ones. A plant tracks light without wanting sunlight. An animal responds to pain without deliberating about it. The appropriate vocabulary for each level of capacity is specific to that level.

Modern ethology extended this precision to animal cognition. De Waal (2016) and Godfrey-Smith (2016) show that animals have genuine perceptual, affective, and cognitive states that can be described accurately using concepts developed for those systems, without importing the full vocabulary of human psychology. This is the model the present paper proposes for AI: not the denial that AI systems do complex and interesting things, but the development of vocabulary adequate to what those systems actually are. The ELIZA effect and its contemporary descendants arise precisely because that vocabulary does not yet exist in public discourse and human psychological terms fill the gap by default.

4. Formal Argument

4.1 Notation

- **C:** phenomenal consciousness — there is something it is like to be this system (Nagel, 1974).
- **I:** intentionality — the system's states are directed toward objects in the world (Searle, 1980).
- **E:** subjective experience — the system undergoes qualitative internal states (Chalmers, 1996).
- **H:** a human psychological attribute: thinking, understanding, intending, or feeling.

4.2 What Human Psychological Attributes Require

- **P_1:** A human psychological attribute H can only be correctly applied to a subject S if S possesses C, I, and E. (Copi & Cohen, 1998; Searle, 1980)
- **P_2:** C requires that there is something it is like to be S — a first-person phenomenal character not reducible to third-person physical or functional description. (Nagel, 1974)
- **P_3:** I requires that S's internal states are genuinely directed toward objects or states of the world, not merely that S produces outputs correlated with such objects. (Searle, 1980)
- **P_4:** E requires that S undergoes qualitative internal states — sensations, perceptions — with a subjective character. (Chalmers, 1996)

4.3 What AI Systems Are

- **P_5:** AI systems are computational processes: sequences of mathematical operations applied to numerical representations of data. The operations are fully describable at the implementation level without reference to any subjective dimension. (Turing, 1950; Russell & Norvig, 2010; Floridi, 2004)
- **P_6:** AI systems manipulate formal symbols according to learned statistical patterns. The outputs of this process are correlated with semantic content because the training data was produced by beings with semantic content. The correlation does not transfer the grounding. (Bender et al., 2021; Searle, 1980)
- **P_7:** There is currently no evidence that any AI system instantiates C, I, or E as defined in P_2–P_4. This claim is empirical and defeasible: it holds given current systems and current evidence. It is not a claim about what future systems might instantiate. (Nagel, 1974; Chalmers, 1996)
- **P_8:** The behavioral level at which AI systems appear to understand or intend is a valid level of description for predicting outputs. It does not license ontological claims about C, I, or E, which belong to a different level of abstraction. (Floridi, 2004, 2008)
- **P_9:** Attributing beliefs and desires to a system because doing so predicts its behavior — the intentional stance (Dennett, 1987) — is a legitimate explanatory strategy. It does not establish that the system ontologically possesses those properties, and it does not provide a basis for assigning causal responsibility or calibrating user trust.

- **P_{10}:** Graded behavioral complexity does not entail human-type consciousness. Entities with sophisticated adaptive behavior — including animals and plants — are not attributed human psychological states in rigorous scientific discourse. The appropriate response to complexity is precise vocabulary, not borrowed psychological terms. (Aristotle, De Anima; de Waal, 1997; Godfrey-Smith, 2016)

4.4 Conclusions

- **C_1:** From P_5–P_7: current AI systems do not instantiate C, I, or E. This is a conclusion argued from the criteria in P_2–P_4 and the empirical description in P_5–P_7. It is not assumed as a premise.
- **C_2:** From P_1 and C_1: applying H to current AI systems is a category mistake — the error of applying a predicate to a subject that does not belong to the category for which that predicate is defined. (Ryle, 1949)
- **C_3:** From P_8–P_9: neither the behavioral plausibility of intentional descriptions nor the predictive utility of the intentional stance overrides C_2. Explanatory validity at the behavioral level does not establish ontological possession at the phenomenological level.
- **C_4:** From P_{10}: the sophistication of AI behavior does not constitute evidence against C_1. Complexity of output is not a marker of phenomenal consciousness.

4.5 Empirical Consequences of the Category Mistake

The following are empirical claims supported by evidence from human-computer interaction and AI ethics research. They are not derived from the logical argument above but are consistent with it and motivate its practical importance.

- **False expectations:** users attribute capabilities to AI systems — genuine understanding, contextual judgment, moral reasoning — that those systems do not have, leading to overtrust and inappropriate reliance. (Waytz et al., 2010; Weizenbaum, 1976)
- **Displaced responsibility:** describing AI as a decision-maker obscures the fact that responsibility for outcomes lies with designers, deployers, and users.
- **Degraded evaluation:** anthropomorphic framing shifts attention from objective performance metrics to subjective impressions of intelligence, making rigorous evaluation harder.
- **Semantic instability:** extending human psychological predicates to AI without redefinition removes their established meaning, creating ambiguity in legal and ethical discourse where precision is required.

5. From Data to AI: Technical Foundations

The following sections build the technical picture of what AI systems actually are. Each section connects to the argument: the goal is to show, at each level of the system, what is happening computationally and what that excludes at the phenomenological level.

5.1 The DIKW Hierarchy

The DIKW hierarchy — Data, Information, Knowledge, Wisdom — describes how raw signals become structured and actionable. A datum is a signal with no inherent meaning: a pixel value, a sensor reading, a character. Information is data organized to provide context. Knowledge is patterns extracted from information that support reliable inference. Wisdom is the application of knowledge in context to achieve a purpose.

AI systems operate at the knowledge level: they extract patterns from large amounts of information and use those patterns to generate outputs. The wisdom level — judgment grounded in values, experience, and contextual understanding — is not instantiated by the optimization process that produces the patterns. This is not a limitation that more data or larger models resolve: it is a structural feature of the approach.

5.2 Databases and Pipelines

AI systems are trained on data stored in relational databases (tables, SQL), NoSQL stores (documents, graphs), Data Warehouses (historical analytics), or Data Lakes (raw unstructured data). The data reaches a model through Extract-Transform-Load pipelines in which cleaning, normalization, and encoding convert real-world signals into numerical arrays. What enters the model is not the world but a numerical representation of a representation of the world. The model never sees words, images, or sounds in any human sense: it processes floating-point numbers. There is no semantic content at the input level, only structure.

6. Programming Paradigms: Declarative vs. Imperative

The contrast between imperative and declarative programming has philosophical implications for how we understand AI agency, or its absence.

The imperative paradigm specifies how a computation is performed: explicit instructions that modify system state step by step. The programmer controls every action. This paradigm most strongly supports the illusion of human-like agency because the programmer's intent is visibly encoded in each instruction.

The declarative paradigm specifies what outcome is desired, leaving the how to the system. TensorFlow computation graphs, SQL queries, and Keras model definitions are declarative: the programmer defines a structure and an objective, and the framework finds a solution. This paradigm is more honest about the nature of AI: the programmer is specifying a mathematical problem and delegating its solution to an optimizer, not issuing instructions in any human sense.

6.1 Code Comparison

Imperative learning loop:

```
for epoch in range(epochs):
    for x, y in data:
        y_pred = model(x)
```

```
loss = (y_pred - y)**2
grad = compute_gradient(loss)
update_weights(model, grad)
```

Declarative equivalent in Keras:

```
model.compile(optimizer='adam', loss='mse')
model.fit(data, labels, epochs=10)
```

The declarative version delegates the computation. The framework computes gradients automatically and updates weights according to the optimizer's rules. There is no agency: no decision, no intent, no goal in the psychological sense. There is a mathematical objective and an algorithm that reduces it.

6.2 Paradigm Comparison

Dimension	Imperative	Declarative
Level of control	Explicit, sequential	Implicit, abstract
Knowledge form	Procedural	Relational
Expression	Step-by-step instructions	Conditions and objectives
Programmer role	Operator	Designer
Use in AI	Kernels, memory, optimizers	Model definition, training objectives

Modern AI systems hybridize both paradigms: the low-level execution engine (C++, CUDA) is imperative; the model definition (Python, TensorFlow, PyTorch) is declarative. The declarative layer formulates the learning problem; the imperative layer solves it on hardware. Neither layer involves intention, understanding, or experience.

7. Machine Learning

Machine Learning is the set of techniques that enables a system to improve its performance on a task by adjusting its parameters based on data, without those adjustments being explicitly programmed. The word learning is a functional metaphor: it describes what the system does from the outside without implying anything about inner process.

7.1 Pipeline

- **Data collection:** transactional databases, sensors, logs, images, text. Labels in supervised learning are numbers or categories, not meanings.
- **Feature engineering:** variable selection, normalization, encoding, dimensionality reduction (PCA, t-SNE). These are numerical transformations.

- **Algorithm selection:** supervised (Decision Trees, SVM, Logistic Regression), unsupervised (K-means, DBSCAN), reinforcement (Q-Learning, DQN). Each maps numerical inputs to numerical outputs through parameter fitting.
- **Training:** minimization of a cost function (cross-entropy, MSE) through Stochastic Gradient Descent, Adam, or RMSProp. Regularization (L1/L2, dropout) constrains the parameter space.
- **Evaluation:** accuracy, recall, F1-score, AUC-ROC. These measure statistical performance, not intelligence.

8. Deep Learning

Deep Learning is a subfield of ML in which the model is a neural network with many layers. The term neural is a metaphor: artificial neurons are mathematical functions that compute a weighted sum of inputs, add a bias, and apply a non-linear activation function. They are not biological neurons.

Stacking layers produces a composite function capable of approximating complex input-output mappings. Specialized architectures include convolutional networks (CNNs) for spatially structured data, recurrent networks (RNNs, LSTMs, GRUs) for sequential data, and Transformers with self-attention for long-range dependencies in language. Each is a mathematical structure chosen for computational efficiency on a class of problems, not for resemblance to how the brain works.

Backpropagation trains the network: the forward pass computes a prediction, the loss function measures error, and the backward pass uses the chain rule to compute gradients with respect to every parameter. The optimizer updates parameters to reduce the loss. After training, the network has parameter values that produce low loss on the training distribution. There is no understanding of the data in this process: there is a mathematical function whose parameters have been adjusted to minimize a scalar. The hierarchical representations that deep networks learn — edges and textures in early layers, semantic abstractions in deep layers — are better described as efficient statistical compression than as conceptual understanding.

9. How a Generative AI Is Built

The following pipeline describes the construction of a large language model. Each stage is presented with its implication for the anthropomorphism question.

- **1. Define purpose and domain:** a design decision made by humans. The system does not form goals; it is given an optimization target.
- **2. Data collection and curation:** the model learns the statistical structure of the training data. It learns nothing about what that data means in human experience.
- **3. Tokenization:** text is converted to integers using Byte-Pair Encoding or WordPiece. The model processes integers, not words.

- **4. Architecture:** a Transformer with self-attention, residual connections, layer normalization, and positional encodings. A mathematical structure.
- **5. Pretraining — next-token prediction:** the model learns to predict the next integer in a sequence. This produces a system that has learned statistical regularities of language. It does not produce a system that knows what words mean.
- **6. Fine-tuning:** human-generated prompt-response pairs adjust outputs toward a desired style. The values in fine-tuning data come from humans, not from the model.
- **7. RLHF:** human raters rank outputs; a reward model learns those preferences; PPO adjusts the generator to maximize reward. Human preferences are externally imposed optimization targets.
- **8. Evaluation and deployment:** statistical testing, compression, containerization, API exposure, continuous monitoring.

10. Hallucinations

The term hallucination, as applied to AI, refers to the generation of text that is syntactically fluent and contextually plausible but factually false. A language model might state that the Paris Opera House opened in 1785 when it opened in 1875. The term is borrowed from the perceptual experience in which a human perceives something that is not there.

The borrowing is misleading. A human hallucination involves a subject who is deceived, an experience of perceiving, and a gap between that experience and the world. None of these elements are present in the AI case. What happens is this: given the preceding token sequence, the model assigns high probability to the token "1785" because the statistical patterns in its training data make that sequence locally plausible. There is no perception, no deception, and no gap between experience and reality, because there is no experience. There is a probability distribution and a sampled token that happens to be false.

Bender et al. (2021) describe this structure clearly: the model has learned to produce text that sounds like text a knowledgeable person would produce, without having access to the knowledge that would ground it in truth. Fluency and factuality are independent dimensions of language model outputs. A system optimized for fluency will produce fluent falsehoods as readily as fluent truths. The practical corrective is not better perception but engineering solutions: retrieval systems, fact-checking pipelines, output verification. These are responses to a statistical problem, not a perceptual one.

11. Anthropomorphization of Language

Anthropomorphization is the attribution of human psychological properties to non-human entities. It is a well-documented cognitive tendency: humans attribute minds to clouds, cars, and simple programs. In the context of AI, this tendency is amplified because language models produce output in human language, which is the most direct external signal of a human mind.

Weizenbaum (1976) documented the mechanism first: ELIZA users attributed understanding to a pattern-matcher because it responded in their language. Waytz et al. (2010) show that people make systematically different trust and moral judgments about systems described in psychological terms versus mechanical terms, even when the underlying system is identical. The vocabulary choice has causal effects on user behavior. This makes precise description not merely an academic preference but a practical requirement for responsible deployment.

The corrective is specific substitutions:

- Replace “the AI thinks that…” with “the model assigns high probability to…”
- Replace “the model understands the context” with “the model produces outputs consistent with the statistical patterns in its training data for this input type.”
- Replace “the AI hallucinated” with “the model generated a false statement — a token sequence that is statistically plausible but factually incorrect.”
- Replace “the AI decided” with “the system produced the output with the highest score under its objective function.”

12. Ethics: Anthropomorphization as an Accountability Problem

The standard AI ethics framework — transparency, fairness, privacy, sustainability — is widely cited but rarely connected to a unified argument. This section connects each principle to the central thesis: anthropomorphization does not merely misdescribe AI systems, it specifically undermines each of these principles at the level of causal mechanism. The argument draws on Floridi's (2004, 2008) observation that causal intervention and accountability must operate at the implementation level of abstraction, not the behavioral level.

12.1 Transparency

Transparency requires that the operation of an AI system be explainable to those it affects. Anthropomorphic language undermines transparency directly because it replaces a causal explanation with a psychological attribution. Saying “the AI decided” or “the model understood” describes the behavioral level while concealing the implementation level at which the actual cause of the output lies: the optimizer converged to parameter values that assign high probability to this output given this input. Floridi's framework makes this precise: transparency requires operating at the implementation level, because that is the level at which the causes of outputs can be identified and modified. Psychological descriptions, however natural, describe the wrong level of abstraction for the purpose of accountability.

12.2 Fairness

Fairness audits examine whether a model produces systematically different outputs for different groups. These audits depend on treating the model as a statistical system whose behavior can be measured and changed. Anthropomorphic framing — attributing differential outputs to the model failing to understand cultural context, or to the model being biased in the way a person is biased — misidentifies the locus of the problem. The model did not fail to understand anything. Its training data contained statistical patterns that produced differential outputs. The fix is at the

implementation level: data reweighting, threshold adjustment, architectural modification. Anthropomorphic framing locates the problem at the wrong level of abstraction and delays the corrective.

12.3 Privacy

Users disclose more sensitive information to systems they perceive as understanding and empathizing with them. Weizenbaum (1976) documented this with ELIZA. Large language models amplify the effect substantially: a user who believes a conversational AI understands them is more likely to disclose medical, financial, and emotional information than one who understands they are interacting with a statistical text generator. Anthropomorphization creates a privacy risk that does not exist when the system is accurately described. Informed consent to data collection and processing requires that users understand what kind of system they are interacting with. When the system is described in psychological terms that imply understanding and judgment, that consent is obtained under a material misrepresentation.

12.4 Sustainability

When AI systems are attributed capabilities they do not have, they are deployed in contexts where their actual performance is insufficient, requiring expensive cycles of retraining, expansion, and iteration. Accurate description of AI capability is a prerequisite for appropriate scoping of AI application, which has direct implications for computational resource use. Additionally, pursuing anthropomorphic goals — general intelligence, understanding, reasoning in the human sense — may drive scale beyond what well-defined functional objectives would require.

12.5 Unanthropomorphization as Ethical Obligation

Describing AI systems accurately is an ethical obligation grounded in the principle of informed consent. Affected users, communities, and regulatory bodies have a right to know what kind of system is acting on them. A system described as intelligent, understanding, or capable of independent judgment generates different expectations and different consent than one described as a statistical optimizer. If the second description is the accurate one, the first is a form of misrepresentation regardless of intent. The obligation to accurate description applies to developers, deployers, journalists, and researchers. It is not satisfied by technical documentation that uses precise language while public-facing communication uses psychological terms.

13. Code: What the System Actually Does

The following two code examples are analyzed as philosophical evidence, not as programming demonstrations. Each component is connected to a specific claim in the argument. The goal is to show, at the level of actual computation, what these systems do and what that excludes.

13.1 Example 1: Variational Autoencoder

A Variational Autoencoder (VAE) learns to compress images into a low-dimensional latent space and reconstruct them. It is a generative model: from a point in the latent space it produces an image. The question the exegesis addresses is what the generative process actually is.

```
# Data loading
(x_train, _), (x_test, _) = mnist.load_data()
x_train = x_train.astype('float32') / 255.
x_train = np.expand_dims(x_train, -1)   # (N, 28, 28, 1)
```

The images are converted to floating-point arrays in [0, 1]. The model never sees digits: it sees matrices of numbers. Normalization makes this explicit — what enters the encoder is not the digit 7 but a 28×28 array of values near 0 (background) or 1 (ink). There is no perception of the digit, only numerical transformation.

```
# Encoder output
z_mean    = layers.Dense(latent_dim, name='z_mean')(x)
z_log_var = layers.Dense(latent_dim, name='z_log_var')(x)
```

The encoder produces two small vectors: a mean and a log-variance in the latent space. This is the system's entire internal representation of an image. There is no concept of what the digit looks like, no recognition in any phenomenological sense. There is a learned mapping from pixel arrays to parameter vectors.

```
# Reparameterization
epsilon = tf.random.normal(shape=(tf.shape(z_mean)[0], latent_dim))
z = z_mean + tf.exp(0.5 * z_log_var) * epsilon
```

This step is frequently described as the model's creativity. It is stochastic sampling from a parameterized probability distribution: a random draw from a Gaussian, shifted and scaled by learned parameters. There is no imagination, no intent, no aesthetic judgment. The apparent creativity is a statistical property of the learned latent space, not a cognitive property of the system.

```
# Loss function
recon_loss = tf.reduce_mean(
    tf.keras.losses.binary_crossentropy(inputs, reconstructed))
kl_loss = -0.5 * tf.reduce_mean(
    1 + z_log_var - tf.square(z_mean) - tf.exp(z_log_var))
self.add_loss(recon_loss + kl_loss)
```

The model's entire objective is encoded in two terms: a reconstruction penalty and a regularization term. Every learned behavior of the system is the product of reducing this scalar over millions of training steps. There is no goal, no motivation, no desire. There is a number to be minimized and an algorithm that minimizes it.

```
# Training
vae.fit(x_train, x_train, epochs=30, batch_size=128,
        validation_data=(x_test, x_test))
```

```
# Generation
z_sample  = np.array([[xi, yi]])
x_decoded = decoder.predict(z_sample, verbose=0)
plt.imshow(figure, cmap='Greys_r')
plt.axis('off')
plt.show()
```

Sampling different points in the latent space produces digit-like images. Moving continuously through the space produces smooth transitions between digit classes. This is accurate at the statistical level: the network has found a compact parametric representation of the statistical structure of its training data. Compression is not comprehension. The system has no concept of number, no knowledge that these shapes carry meaning in human symbolic systems.

13.2 Example 2: Next-Token Prediction and Hallucination

A language model generates text by repeatedly predicting the most probable next token given a context. The following example shows how hallucination occurs as a feature of probability-based generation, not as a malfunction of perception.

```
def generate(model, tokenizer, prompt, max_tokens=60, temperature=1.0):
    input_ids = tokenizer.encode(prompt, return_tensors='pt')
    generated = input_ids.clone()

    for _ in range(max_tokens):
        with torch.no_grad():
            outputs = model(generated)
            # logits over vocabulary for the next token position
            logits = outputs.logits[:, -1, :]

        # Temperature scales the distribution
        # high T = more uniform = more surprising and more often false
        # low  T = more peaked  = more conservative
        logits = logits / temperature
        probs  = F.softmax(logits, dim=-1)

        # This is not a decision. It is a weighted random draw.
        next_token = torch.multinomial(probs, num_samples=1)
        generated  = torch.cat([generated, next_token], dim=-1)

        if next_token.item() == tokenizer.eos_token_id:
            break

    return tokenizer.decode(generated[0], skip_special_tokens=True)
```

Every step is explicit. The model maps an integer sequence to a probability distribution over the vocabulary. A token is sampled from that distribution. The token is appended and the process repeats. There is no comprehension at any step.

When this process produces “The Paris Opera House opened in 1785,” what has happened is: given the preceding sequence, the model assigned higher probability to ‘1785’ than to ‘1875’, because the statistical patterns in its training data made that assignment. The model did not misremember, did not hallucinate in any perceptual sense, and did not intend to mislead. It sampled from a distribution. The sample was false.

The temperature parameter makes the mechanism visible: at high temperature the distribution is nearly uniform and the model generates more varied, more frequently false text. At low temperature it concentrates on the most probable tokens and generates more conservative text. No epistemic state is being modulated. A softmax parameter is being changed. This is what the term hallucination describes when applied to a language model: a statistical property of the generation process, not a perceptual malfunction.

14. Conclusion

This paper has argued, through conceptual analysis, philosophical engagement, formal logic, and code exegesis, that applying human psychological vocabulary to current AI systems is a category mistake with specific measurable consequences. The formal argument establishes that human psychological attributes require phenomenal consciousness, genuine intentionality, and subjective experience — properties for which there is no current evidence in AI systems. This conclusion is reached by arguing from established criteria rather than assuming it, and by engaging the principal objections directly.

Dennett's intentional stance is acknowledged as a valid predictive tool while its limits as a basis for ethical accountability are identified. Hinton's observations about large-scale neural networks are taken seriously and used to support the case for new precise vocabulary rather than expanded anthropomorphism. Searle's Chinese Room is used diagnostically rather than as a foundation, with the primary weight of the argument resting on Nagel's phenomenological criterion and Floridi's levels of abstraction. The Aristotelian parallel and de Waal's concept of anthropodenial together clarify that the paper's position is not a blanket refusal to attribute properties to AI systems — it is a call for vocabulary adequate to what those systems actually are.

The ethics section connects each standard AI ethics principle to the anthropomorphism problem at the level of causal mechanism, drawing on Floridi's framework to argue that accountability requires operating at the implementation level of abstraction rather than the behavioral level. The code exegesis makes concrete what abstract philosophical claims assert: the VAE's apparent creativity is stochastic sampling; the language model's hallucination is a high-probability false token; the optimizer's apparent goal is a scalar to be reduced.

Several questions remain open and are identified here for future work. First, this paper calls for AI-specific vocabulary but does not propose a systematic framework for it. Developing such a vocabulary — drawing on existing technical terminology in ML, cognitive science, and information theory — is a necessary next step. Second, the paper's argument is bounded by

current systems and current evidence. Criteria for what would constitute credible evidence of phenomenal consciousness in an AI system — criteria that are neither trivially easy nor impossibly demanding to satisfy — need to be developed. Third, the practical question of how precise technical vocabulary can displace psychological terms in public and media discourse, given the cognitive tendency toward anthropomorphization that Weizenbaum identified, is an open problem in science communication that this paper does not resolve.

References

Aristotle. De Anima [On the Soul]. (J. A. Smith, Trans.). In W. D. Ross (Ed.), The Works of Aristotle (Vol. 3). Clarendon Press. (Original work composed ca. 350 BCE)

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford University Press.

Copi, I. M., & Cohen, C. (1998). Introduction to logic (10th ed.). Prentice Hall.

de Waal, F. (1997). Are we in anthropodenial? Discover, 18(7), 50–53.

de Waal, F. (2016). Are we smart enough to know how smart animals are? W. W. Norton.

Dennett, D. C. (1987). The intentional stance. MIT Press.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Floridi, L. (2004). Levels of abstraction and the semiotics of emergent meanings. In C. Murata (Ed.), The Semiotics of Cellular Communication in the Immune System. Springer.

Floridi, L. (2008). The method of levels of abstraction. Minds and Machines, 18(3), 303–329. https://doi.org/10.1007/s11023-008-9113-7

Godfrey-Smith, P. (2016). Other minds: The octopus, the sea, and the deep origins of consciousness. Farrar, Straus and Giroux.

Hinton, G. E., Bengio, Y., & LeCun, Y. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Marcus, G., & Davis, E. (2019). Rebooting AI: Building artificial intelligence we can trust. Pantheon Books.

McCarthy, J. (1956). Proposal for the Dartmouth summer research project on artificial intelligence. Reprinted in AI Magazine, 27(4), 12–14 (2006).

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115–133. https://doi.org/10.1007/BF02478259

Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83(4), 435–450. https://doi.org/10.2307/2183914

Russell, S., & Norvig, P. (2010). Artificial intelligence: A modern approach (3rd ed.). Prentice Hall.

Ryle, G. (1949). The concept of mind. Hutchinson.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (2021). Evidenced-based fact-checking. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 1–10). Association for Computational Linguistics.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Waytz, A., Cacioppo, J. T., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. Perspectives on Psychological Science, 5(3), 219–232. https://doi.org/10.1177/1745691610369336

Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation. W. H. Freeman.