ChatGPT

AI Epistemology: Foundations, Limits, and Handling the "Unknown Unknowns"

Introduction

As AI systems rapidly permeate every sector – from finance and law to healthcare and education – they are reshaping the very conditions of our **epistemic enterprise** [1]. This raises a fundamental question: do AI outputs constitute genuine **knowledge**, or merely the appearance of knowledge? Understanding the nature and limits of AI "knowledge" is not just academic; it is critical for **safely deploying AI in real-world, high-stakes applications**. The field of *AI epistemology* examines how AI systems acquire, represent, and use knowledge, and how this compares to human knowing. It overlaps with classical epistemology (the philosophy of knowledge) but also confronts practical constraints: AI must deal with what is knowable given **finite data and computation**, unlike idealized human knowers [2]. In this report, we delve into: (1) the **philosophical foundations** of AI knowledge, (2) the **limits of AI understanding**, and (3) how AI handles **uncertainty and novelty** – the so-called "*unknown unknowns*." We also highlight key thinkers and debates, from early AI pioneers to contemporary AI ethicists, to trace how perspectives have evolved through to **2024–2025**. The goal is a comprehensive review and conceptual analysis that informs both academic insight and practical considerations (e.g. chain-of-thought techniques, critical system design, AI governance, and alignment).

Philosophical Foundations of AI Knowledge

At its core, *AI epistemology* asks: **How do AI systems represent and generate knowledge?** This question has deep roots in both philosophy and early AI research. Classical epistemology defines knowledge as *justified true belief*, but AI brings new twists: machines have no consciousness or beliefs in the human sense, yet they manipulate information in ways that can produce knowledge *for* humans [3]. We can consider AI knowledge from several angles:

- **Knowledge Representation (Symbolic vs. Subsymbolic):** Early AI (1950s–1980s) was dominated by the **symbolic paradigm**, which assumed that intelligence comes from manipulating explicit symbols and rules about the world. Allen Newell and Herbert Simon's *Physical Symbol System Hypothesis* (1976) epitomized this view: *"A physical symbol system has the necessary and sufficient means for general intelligent action."* [4] In practice, this led to *knowledge-based systems* using logic, ontologies, and if-then rules to encode what the AI "knows." John McCarthy (who coined the term AI) argued that AI must confront *epistemological problems* distinct from philosophy: philosophers imagine an ideal reasoner with unlimited observation and computation, whereas AI must figure out what is knowable with **limited data and processing** [2]. Classic AI formalisms (first-order logic, frames, semantic networks, etc.) sought to represent facts about the world and reason over them. However, these formalisms often proved **"epistemologically inadequate for general intelligence"**, as McCarthy admitted – they struggled to achieve broad competence even given unlimited speed [5] [6].

- **Implicit Knowledge and Machine Learning:** By the 1990s and especially the 2010s, the rise of *machine learning* (connectionist paradigms) shifted the focus to **subsymbolic representations**. Instead of explicit facts, AI "knowledge" became encoded in the parameters of neural networks trained on data. Modern large-scale AI, like deep neural networks and especially large language models (LLMs), encapsulate knowledge as complex **statistical patterns**. For example, GPT-style LLMs compress vast text corpora into high-dimensional weight space, such that they can generate coherent answers. This aligns with the view of knowledge as **compressed information**: *"knowledge is encapsulated in embedded vectors of patterns learned from data… an optimized, compressed representation of sensory inputs that retains essential information"* [7] . In other words, machines "know" inasmuch as their models encode useful regularities (e.g. an LLM "knows" that Paris is the capital of France because this fact was embedded in its training text). This sub-symbolic approach overcame some brittleness of rule-based AI by generalizing from examples, but it introduced new epistemic challenges (like **opacity** and lack of explicit reasoning, discussed later). Notably, earlier skeptics like Hubert Dreyfus, who emphasized the importance of non-symbolic, intuitive knowledge, were vindicated when neural networks succeeded; human expertise indeed relies on *"fast intuitive judgments rather than step-by-step symbolic reasoning,"* a point Dreyfus made as early as the 1960s [8] (anticipating modern insights about intuitive vs. analytical cognition [9]). Today's best AI systems blend both approaches: for instance, a self-driving car might use logical rules for traffic laws (symbolic knowledge) alongside neural nets for vision (learned implicit knowledge).

- **Do AI Systems** Know **Anything?** From a strict philosophical standpoint, whether AI *truly "knows"* is debatable. AI lacks **conscious belief** or understanding; it manipulates symbols or numbers without subjective awareness. As one 2025 analysis succinctly puts it, *"AI systems do not know in the human, doxastic sense (no beliefs, no accountability)"* [3] . They don't possess **justified true belief** in the way a human scientist does, because there is no conscious agent having insights. However, AI can still be a **source of knowledge** *for us*. Ratiomachina (2025) describes three ways AI contributes to human knowledge without "believing" anything itself [3] : **(1)** as an *"epistemic instrument"* that extends our cognitive capacities (much like a microscope reveals truths we can record, an AI can simulate scenarios or calculate results we learn from), **(2)** via *"artificial testimony"* – an AI's outputs can inform and justify human beliefs if the system is sufficiently *reliable* (more on reliability below), and **(3)** by *"scaffolding understanding"* – if we can interpret an AI's internal mechanisms or explanations, it can help us understand *why* something is true [10] [11] . In short, while an AI doesn't **know** in the way humans do (it has no conscious grasp or accountability), it can *produce* knowledge that humans leverage.

- **Justification and Trust in AI Outputs:** A crucial aspect of knowledge is justification – why should we trust a purported fact or prediction? Traditional epistemology demands reasons or evidence. With AI, this leads to debates on **transparency vs. performance**. Modern AI systems, especially deep learning models, are often *"essentially opaque — no single agent (not even the developers) can survey all the epistemically relevant steps"* in how the AI arrives at an answer [12] . One might think that without transparency, we cannot trust the AI's "knowledge." However, emerging views in philosophy of AI suggest an alternative: **computational reliabilism** [12] . This approach justifies using AI outputs not by peering into every weight and neuron, but by *empirically* demonstrating that the overall system is **reliably truth-conducive**. In other words, if an AI method can be shown through rigorous testing, validation, and error-checking to consistently produce accurate results, we can treat its outputs as *knowledge* even if the internal process is opaque [13] . For example, we might not interpret every parameter of a climate simulation, but we trust it because it's validated against known data and physical laws (*analogy:* we don't inspect every transistor of a calculator to trust that $2+2=4$; we trust it because of reliable design). Thus, much recent work on AI alignment and safety emphasizes **process-oriented validation**: benchmark performance, calibration of confidence, adversarial testing, and uncertainty quantification, rather than full transparency [13]. This shift acknowledges that *demanding full interpretability would forfeit many benefits of complex AI* (which often surpass human cognitive capacity) [12]. Instead, we seek a **warrant for trusting AI** akin to how we trust instruments: by their proven reliability under well-designed oversight regimes.

- **AI as Cognitive Tools – Extended Mind:** Philosophically, we can view AI systems as part of an **extended epistemic infrastructure** for humanity. They are **cognitive artifacts** that, like writing or telescopes, augment human knowledge processes [14]. On this view, it matters less whether the *machine itself* "understands," and more whether humans, with the machine's aid, can achieve new understanding. Indeed, AI's greatest epistemological impact may be in enabling **new forms of knowledge**. For instance, complex simulations and machine learning can detect patterns no human would notice unaided. This is reminiscent of philosopher Paul Humphreys' idea of the **epistemology of computational science** – simulations can yield novel insights but are opaque, shifting humans to more of a supervisory role [15]. AI systems have already demonstrated the ability to find **"unknown knowns"** – latent truths hidden in data – which humans then interpret. (One example: an ML model might reveal a subtle statistical pattern in medical data that points researchers to a new hypothesis.) This dynamic has led some to argue that AI can generate *"epistemic novelty"* for us even if it isn't a "knower" itself [11]. In summary, the foundation of AI knowledge is a blend of **formal philosophy, cognitive science, and engineering**: it forces us to refine what we mean by knowing and to distinguish **knowledge *in* AI vs. knowledge *through* AI** [11]. As a 2025 Medium essay advises, we should treat advanced AI **"primarily as cognitive artifacts and reliable instruments"** for extending human knowledge, rather than anthropomorphizing them as independent knowers [16]. This perspective grounds our use of AI in a way that *strengthens* the epistemic foundations of society rather than undermining them [1].

Limits of AI Understanding

Even as AI systems grow more capable, there are fundamental questions (and critiques) about whether they truly **understand** what they process, and what inherent limits might block AI from reaching human-level (or beyond) understanding. Key debates and theoretical limits include:

- **Syntax vs. Semantics – The Chinese Room:** Philosopher John Searle's famous *Chinese Room* argument (1980) posits a clear limit: a computer following a program (manipulating symbols based on formal rules) **lacks understanding** of those symbols. Searle asked us to imagine a person in a room following instructions to manipulate Chinese characters. To an outside observer, it might appear the room **speaks Chinese**, but Searle argued that neither the person nor the computer program *actually knows* Chinese – there is only symbol processing without semantics [17] [18]. His conclusion was that *"as long as the program is defined in terms of formal operations on symbols, it cannot have meaningful understanding."* [18] In other words, *syntax is not semantics*. This thought experiment has been a touchstone in AI philosophy, suggesting a hard limit: *No purely computational system, just by executing rules, can generate real understanding or consciousness*. However, this argument is not settled. Recent developments with fluent LLMs like GPT-4 have prompted some to **re-examine Searle's claim**. Some researchers argue that while LLMs might not prove true understanding, they *undermine the intuition* that mere rule-following can't *appear* understanding. For example, Siemers (2025) notes that modern AI systems produce outputs *so fluent and contextually appropriate* that the classical Chinese Room intuition ("it's just a mindless rulebook") loses force – the *"mere rule-following"* dismissal *"no longer feels like a sufficient explanation for the rich responses"* of systems like GPT-4 [19]. Nonetheless, even these commentators usually stop short of claiming the AI **truly understands**; they instead suggest we need new frameworks for what machine understanding would entail [20]. Some responses to Searle also point out potential solutions: the **Robot Reply**, for instance, says that if the program were embodied in a robot with cameras and limbs, connected to the world, it *could* ground the symbols in sensory experience and thus achieve understanding [21]. Hans Moravec famously wrote: *"If we graft a robot to a reasoning program, we wouldn't need a person to provide the meaning anymore – it would come from the physical world."* [21] This implies that lack of **embodiment** and real-world causal interaction is what kept the Chinese Room ignorant. Indeed, modern AI research in robotics and embodied agents follows this intuition, aiming to ground AI in sensorimotor experience so that its "symbols" refer to real objects and not just other symbols.

- **Tacit Knowledge and Intuition:** A related limitation is the challenge of **tacit knowledge** – the kind of intuitive know-how that humans acquire through experience but have trouble explicitly articulating (originally highlighted by philosopher Michael Polanyi as "we know more than we can tell"). AI critics in the 1960s–70s, notably Hubert Dreyfus, argued that human intelligence *"depends primarily on fast, intuitive judgments rather than explicit symbolic manipulation."* [8] Early AI failed dramatically at tasks requiring commonsense or embodied skills (e.g. robot navigation, commonsense reasoning) because so much of human knowledge is *subconscious or context-bound*. Dreyfus pointed out that human experts do not follow millions of formal rules in real time; rather, they develop gut feelings, pattern recognition, and inarticulate skills. He doubted that formal rules or logic could ever capture this rich, situation-specific know-how. This critique embarrassed AI researchers in the 1970s (Dreyfus' book *What Computers Can't Do* was controversial), but over time many of his points were assimilated. The field of **expert systems** in the 1980s tried to encode experts' tacit knowledge as rules, but often fell short. Today, deep learning can capture some tacit patterns (like intuition in Go or chess moves) by training on vast experience, yet it still struggles with the breadth of human commonsense. Notably, cognitive science eventually paralleled Dreyfus: Daniel Kahneman's dual-process theory (System 1 vs System 2) basically affirmed there's a fast, intuitive mode of cognition that isn't easily reducible to explicit reasoning [9]. Modern AI has made progress – e.g. reinforcement learning can train robots with *trial-and-error* to acquire skills that were hard to hand-code – but broad human-like intuition in open environments remains an unsolved frontier. The **Frame Problem** in AI (first noted by McCarthy & Hayes) exemplifies this difficulty: it's the challenge of telling an AI which facts *not* to worry about in a changing world (humans intuitively know that many things remain irrelevant or unchanged when an action is taken, without checking every fact; computers do not know what can be ignored without explicit programming). Despite many attempts, a general solution to the frame problem – essentially, how to imbue machines with *common sense relevance* – has proven elusive and is considered a fundamental limit by some philosophers of AI.

- **Statistical Learning and "Understanding":** The recent success of large language models reignites the question of whether *statistical correlation at scale is closing in on genuine understanding*. Critics like Emily M. Bender argue that current LLMs are **"stochastic parrots"** – they adeptly regurgitate patterns in language without any true grasp of meaning or truth [22]. The *"stochastic parrot"* metaphor, coined in 2021, warns that we shouldn't mistake an LLM's fluent output for actual comprehension: the model *"stitches together sequences of linguistic forms"* based on probability, with *“no genuine meaning at all behind the text”* [22]. Supporting this, LLMs notoriously produce **hallucinations** (plausible-sounding falsehoods) which a true understanding agent would avoid. On the other hand, some AI researchers contend that as models incorporate more data and perhaps neural architectures evolve, they *do* start to exhibit glimmers of understanding – at least in a behavioral sense. For instance, GPT-4 can solve problems that seem to require multi-step reasoning or analogies, leading to arguments that it has a kind of “emergent” model of the world. This debate is ongoing. It touches on whether **meaning** can arise from pure data regularities or whether something fundamental (like grounding in the real world, or a model of **causality**) is missing. Judea Pearl, a pioneer of Bayesian networks and causal inference, has been a vocal critic of purely statistical AI. He argues that *current AI “learns associations” but fails at understanding cause and effect*, which is critical for true reasoning and generalizing beyond observed patterns [23]. In Pearl’s view, without causal models, AI will always be stuck at a superficial understanding level – correlation, not *causation*. His **Ladder of Causation** concept suggests that human-like intelligence requires mastering interventions and counterfactuals (“What if…?” reasoning), something current neural networks struggle with unless explicitly trained for. Recent work in AI is indeed trying to integrate causal reasoning and **world models** to overcome these limits.

- **The No Free Lunch Theorem:** There are also **theoretical limits** on what any AI (or any algorithm) can do. A famous result in machine learning is the *“No Free Lunch” (NFL) theorem*, which states that *no single learning algorithm can outperform all others on every possible problem* [24]. In simplified terms, if an AI is tuned to excel in one class of environments or tasks, there will exist some other environment where it performs worse than an alternative. This has epistemological implications: there is no **universal problem-solver** or omniscient AI that is uniformly best at everything. **Inductive bias** (assumptions built into a model) is always needed, and that bias will suit some domains and not others. For AI, this means any knowledge a system has is *particular* to the data and world it has experienced or encoded. If it ventures far outside that distribution, it can fail. Thus, there will always be “blind spots” or weaknesses – an inherent limit on *general* understanding. (Though techniques like meta-learning and massive scaling try to make AI as general as possible, NFL implies trade-offs will remain.) Another theoretical limit often cited is Gödel’s Incompleteness Theorem, which some (like Roger Penrose) interpreted to mean human mathematicians can see truths no formal system can prove. However, most experts find that argument unconvincing as a practical limit on AI (it applies to formal logical systems and doesn’t straightforwardly translate to real-world reasoning). Still, it reminds us that in any sufficiently powerful system, there may be true statements it cannot prove within its own framework – an abstract limitation.

- **Transparency and Interpretability Limits:** Understanding is also limited by our ability to **interpret** the AI. Even if an AI had correct knowledge, if it cannot explain its reasoning or content, we might say it doesn’t “understand” in a meaningful way (or at least, *we* cannot understand what it “understands”). This is the problem of **epistemic opacity** in AI. As mentioned earlier, one solution is to rely on reliable performance, but for critical decisions, stakeholders often demand interpretable reasons. Current state-of-the-art models like deep networks are complex webs of numbers that do not readily translate to human-comprehensible concepts. Research in **explainable AI (XAI)** and **mechanistic interpretability** tries to open this black box. For example, **mechanistic interpretability** seeks to map parts of a neural network to human-meaningful functions (e.g. this neuron detects cars, this layer encodes grammatical structure) [25]. If successful, this could bridge machine processing to human-style understanding. However, interpretability itself might face a scalability limit: as models get more complex, fully understanding their internals may be as hard as understanding a complex natural system. There's a Catch-22: the more an AI *learns* and internalizes, the harder it may be to distill that into simpler explanations. This raises the question of whether *the concept of "understanding" is being partially redefined* by AI – perhaps we will accept a form of understanding that is distributed and inextricable from big data, which no single mind can grasp, only *tools* can verify (as long as the results align with reality). This ties back to the earlier point that **prediction isn't the same as understanding**, but with interpretability advances, we hope to *connect* the two [25] .

In sum, current AI systems **excel at certain facets of "knowing"** (memorizing facts, correlating patterns, optimizing defined goals) yet **lack others** (grounded semantics, commonsense context, self-reflection, and broad adaptability). They don't **know that they don't know** – a crucial aspect of human epistemology (Socratic ignorance) that machines struggle with. Many of these limits are active research areas, and some are philosophical standoffs. Whether future AI can overcome them – achieving something like **true understanding** or consciousness – remains uncertain. Some thinkers believe new paradigms (e.g. combining symbolic and neural methods, or incorporating embodiment and causality) might get us there, while others maintain that machines will at best simulate understanding. Notably, as AI systems approach human-level language and problem-solving, the **line between simulation and reality of understanding** becomes an ever sharper debate in AI alignment and safety communities, since an AI that *seems* to understand people could be very persuasive (whether or not it actually grasps meaning or ethics). This is why careful governance is needed even *before* machines cross the threshold into genuine understanding.

Handling Uncertainty and Novelty (The "Unknown Unknowns")

One of the greatest practical challenges for AI – with deep philosophical implications – is how to deal with **uncertainty, the unknown, and the novel**. In everyday human life, we constantly navigate **known knowns** (things we know), **known unknowns** (things we know we don't know), and we occasionally encounter **unknown unknowns** – surprises we didn't even realize were possible. AI systems, being bound by their programming and training data, historically handle uncertainty in limited ways. Ensuring AI can gracefully handle novelty and flag its own ignorance is crucial for reliability, especially in critical applications (*we don't want a medical AI to confidently assert a wrong diagnosis on an unprecedented symptom*). Here we review how AI manages uncertainty, and recent advances toward grappling with the *"unknown unknowns."*

- **Types of Uncertainty – Aleatoric vs. Epistemic:** In the language of statistics, there are two broad types of uncertainty. **Aleatoric uncertainty** refers to inherent randomness in the world (e.g. the outcome of a fair die roll – no matter how much data you gather, there's irreducible uncertainty). **Epistemic uncertainty** refers to uncertainty due to lack of knowledge – it *can* be reduced if you obtain more information (e.g. uncertainty about a poorly known medical condition can lessen as research is done) [26] . AI systems need to account for both. Classical AI handled uncertainty via probabilistic reasoning. Pioneering work by Judea Pearl in the 1980s introduced **Bayesian networks**, allowing AI to represent uncertain knowledge and update beliefs with evidence. For instance, an AI diagnosis system might assign probabilities to diseases and update them as symptoms are observed. Probabilistic AI provides a calculus for **known unknowns** – if the AI knows the possible hypotheses, it can quantify its uncertainty among them. What's harder is **unknown unknowns**: situations the system never anticipated at all. A Bayesian network can't have a node for a disease it doesn't even conceive of. This is where **open-world** or **open-set** reasoning comes in (more below). Some AI frameworks (like **Dempster–Shafer theory** or **fuzzy logic**) were developed to handle uncertainty in a more flexible or bounded-rationality way than strict probability, acknowledging the ambiguity in inputs and rules. These saw use in expert systems and control systems (fuzzy logic especially in engineering control where inputs are noisy and vague). Today, interest in uncertainty quantification is resurging with deep learning: researchers distinguish between *model uncertainty* (not being sure due to lack of training data – an epistemic uncertainty) and *data noise* (aleatoric uncertainty). Techniques like **Monte Carlo dropout** or **deep ensembles** treat neural networks in a Bayesian manner to estimate confidence in predictions. The key goal is to have AI **not only provide an answer, but also a confidence or an "I'm not sure" when appropriate**.

- **The Problem of Overconfidence and Hallucinations:** A major issue with current AI models, especially large language models, is that they are often **overconfident** and will *guess* or "hallucinate" an answer even when they have no knowledge. This is essentially an epistemic failure – the AI doesn't *know that it doesn't know*. Recent research by OpenAI and others has shed light on why this happens in LLMs. During training, these models are rewarded for producing plausible answers to all questions, never for saying "I don't know." As one analysis explains, *the language model is trained to always predict a next word, even if the query is unlike anything seen before; moreover, not answering is penalized in typical benchmark-oriented training, so the model learns that guessing often yields better aggregate accuracy than abstaining (due to occasional lucky guesses).* [27] In other words, the training process implicitly taught LLMs that *making something up* is better than saying nothing. This is why an LLM will concoct a bibliographic citation or a fictional fact with perfect fluency – it was never properly taught the option to *remain silent or admit ignorance*. To address this, researchers have proposed adjustments to training and evaluation: for example, *don't penalize the model for saying "I don't know" or expressing uncertainty* in certain contexts [28]. If you modify the objectives so that *truthfulness* is rewarded over mere accuracy on known answers, the model can learn to refrain from guessing. Indeed, a 2023 OpenAI paper found that by allowing a model to opt-out of answering when uncertain, you can significantly reduce hallucinations [29]. The ideal outcome is an AI that *"doesn't lie, favoring confident ignorance (e.g., 'I don't know') over bluffing"* [30]. Achieving this is a current priority: imagine a chatbot that, when faced with a question outside its knowledge, reliably responds with caution or a request for clarification instead of a fabricated answer – this would greatly enhance trust.

- **Out-of-Distribution (Novel) Inputs:** No matter how much data we train on, the real world has surprises. An AI might encounter queries or situations that are **out-of-distribution (OOD)** – fundamentally different from its training examples. A robust AI needs to detect this and respond appropriately (e.g. raise an alert, consult a human, or at least not be overconfident). Researchers have been developing **OOD detection** methods: basically, ways for a model to sense *"this input is unlike anything I've seen, so I should be careful."* For classification tasks, one approach is measuring whether an input falls far from all known class clusters in feature space. For language models, it's trickier, but some methods use likelihoods or contradictions among multiple prompts to flag uncertainties. A fascinating illustration of the OOD challenge was given in a recent article [31] [32]: Many LLMs have seen a common riddle about a surgeon, the father and son accident, etc., so they *expect* the answer that the surgeon is the mother. If you twist the riddle's details (making the mother the one who died) – something not common in training – the model can break and give absurd answers because it misinterprets the situation (it tries to force the pattern it "knows"). The AI had **never seen that exact scenario** and thus **hallucinates an answer** (like "the surgeon is the boy's other mother") which is obviously wrong [32] [33]. This shows that when faced with **unknown unknowns**, the model doesn't realize the situation is novel; it shoehorns it into nearest-known patterns. As the author of that analysis put it, *"even if you teach an LM to say 'I don't know,' it can't know* that it doesn't know *what it doesn't know."* [34] In other words, the AI lacks a **meta-knowledge** of its own blind spots. Addressing this remains a hard problem.

- **Techniques for Uncertainty Awareness:** A number of strategies are emerging to make AI more uncertainty-aware and resilient to novelty. One is **Ensemble Modeling** – run multiple models or multiple variations of input (like prompting an LLM multiple times or with different phrasings) and see if they all agree. If they diverge, that's a sign of uncertainty. For example, *chain-of-thought prompting* combined with *self-consistency* will sample several reasoning paths for a question and check if the conclusions converge [35] . If not, the AI (or the system around it) can signal low confidence. Another approach is **conformal prediction and abstention**: designing the AI to output a prediction with a confidence interval, or a prediction set, and if the confidence is below a threshold, output "I abstain." In fact, one paper developed a *"conformal abstention"* method giving guarantees on the rate of incorrect answers vs. abstentions [36] . All these are ways to quantify not just "What does the model predict?" but "How sure is it?".

- **Handling the *Truly* Unknown – Open-World AI:** The frontier of this topic is building AI that can handle **open-world scenarios**, where new classes, new rules, or entirely unexpected phenomena can occur. A subfield called **open set recognition** and **open-world learning** directly tackles this [37] . As an AAAI survey put it, *"no amount of training will prepare the system for all unknown inputs."* [38] We can't just feed the model every possibility, because the space of the unknown is (theoretically) infinite. Thus, the AI must be able to **recognize when something is novel and adapt**. Approaches include: detecting anomalies (data points with low confidence or that activate neurons in unusual ways), maintaining a reject option (if input looks unlike training data, refuse to classify), and even **on-the-fly learning** (incorporating new inputs as new knowledge). For example, imagine an AI that encounters a new type of object for the first time – instead of pigeonholing it into the closest known category, it flags it as "unknown" and perhaps queries a human or database for more information. Some progress has been made: there are deep learning models that incorporate an **OpenMax** layer (an extension of softmax that reserves probability for "none of the known classes") [39] . There are algorithms that explicitly model a **background class** or use distance thresholds in feature space to reject outliers [40] . However, a challenge remains: *unknowns are often mislabeled by models with high confidence* – the model might be **wrong but certain**, which is dangerous [41] . Adversarial examples exploit this by finding inputs that a model confidently misclassifies. The survey noted that for deep networks, novel inputs often map to *some* known-looking activation pattern, making them hard to spot [42] . In short, unknown unknowns can hide in the shadows of known patterns.

Researchers stress that dealing with true novelty is *essential* for robust AI. It requires a balance: *"maintaining accuracy on the core problem while handling the unknown unknowns."* [43] This might involve more **meta-learning** (AI that can learn how to learn, adapting to new tasks), **generative modeling** (to simulate possible novel instances), or incorporating **world knowledge and simulation** so the AI has broader context of what's possible. We also see the use of AI **agents that can do exploratory actions** – for instance, if a query is novel, a chatbot agent might automatically perform a web search or ask clarifying questions, thus reducing its uncertainty by gathering information (a bit like how a human would research an unfamiliar topic). Another idea is **Knightian uncertainty** handling: this is uncertainty that cannot be quantified probabilistically (named after economist Frank Knight). Some alignment researchers propose frameworks like *“infra-Bayesianism”* to handle this kind of uncertainty in AI planning, allowing for model ignorance in a principled way.

- **AI and Creativity – Can AI discover the Unknown Unknowns?** A profound aspect of human intelligence is creativity: the ability to come up with entirely new hypotheses or artifacts. Humans occasionally venture into the “wilds of the unknown” (to quote mathematician du Sautoy) and bring back something truly novel. Can AI do the same, or is it fundamentally a **re-combinator of existing knowledge**? Using the Rumsfeld matrix terms, we might ask: can AI ever *intentionally* explore **Unknown Unknowns**, or is it confined to Known Knowns and Unknown Knowns at best? Lars Godejord (2025) applied the **Known–Unknowns matrix** to AI’s creative potential [44] [45]. In his analysis, today’s AI excels in the **“Unknown Knowns”** quadrant – *surfacing latent patterns or insights implicit in data that humans hadn’t noticed* [46]. For example, an AI might identify a hidden correlation in a scientific dataset (a pattern that was *unknown to us but actually present* – thus an “unknown known”). AI is very good at these because of its brute-force pattern-matching. However, in the **“Unknown Unknowns”** quadrant – *the truly unforeseeable innovations or paradigm shifts* – current AIs fall short [47]. They lack the **intentionality and broad world understanding** to define new problems or leap outside the training distribution on their own. Godejord notes that achievements like discovering penicillin or formulating relativity were not simply extrapolations of existing data; they required framing new questions and imaginative insight [47]. Present AI systems, which operate by recognizing patterns in provided data, *“typically lack the intrinsic drive or deep contextual understanding to define entirely new problems or paradigms without substantial human guidance.”* [47] In practice, this means AI’s “creativity” is often a reflection of human creativity contained in its data or goals. However, this might evolve: techniques like **generative adversarial exploration** or evolutionary algorithms allow AIs to *surprise us* in constrained ways (e.g. an AI can generate a new art style by optimizing for novelty, or come up with odd designs in engineering through evolutionary strategies). Some researchers have even built agents with *curiosity-driven learning*, where the AI is intrinsically rewarded for encountering novel states (used in certain reinforcement learning tasks). Such mechanisms aim to mimic the human urge to explore the unknown. It’s an open question whether scaling these up could lead to machine-driven discovery of unknown unknowns. The consensus for now is that **human-AI collaboration** is the safest bet: AI can illuminate the **terra incognita** by crunching data and proposing patterns, and humans provide the judgment, values, and big-picture framing to navigate which unknowns to explore.

- **Critical Applications and Epistemic Assurance:** In high-stakes or **governance** contexts, handling uncertainty isn’t just a nice-to-have, it’s mandatory. For example, in medical diagnosis AI or self-driving cars or legal advisory systems, a **false certainty** can be fatal or unjust. Thus, practitioners emphasize **confidence calibration** and failsafe mechanisms. One recommended practice is to always have a *“human in the loop”* when the AI’s confidence is below a threshold or when novel cases are encountered. There’s also a push for AI systems to communicate uncertainty in a user-friendly way (for instance, showing a probability or an error bar, or expressing in natural language “I’m not entirely sure about this answer”). Additionally, organizations deploying AI are encouraged to do **stress testing**: deliberately feed edge-case scenarios, adversarial inputs, and rare situations to identify where the AI’s knowledge breaks down [48]. As a recent guideline noted, we should implement *“uncertainty displays, domain-specific grounding, and out-of-distribution guards”* for AI that provides answers [49]. This might mean connecting the AI to a trusted knowledge base for factual queries (so it doesn’t hallucinate facts) and having a monitor on input distribution shift. In AI **alignment** (ensuring AI’s goals and actions remain in line with human values), uncertainty also plays a pivotal role. Notably, Stuart Russell's **principles for beneficial AI** state that *the AI should never be overly confident in the objectives it is given* – it should *always allow for the possibility that its understanding of the human's true preferences is incomplete or incorrect* [50]. This *"machine is initially uncertain about the human preferences"* is Principle #2 in Russell's book *Human Compatible* [51]. The idea is that if the AI is uncertain about the goal, it will be willing to listen, to ask permission, or to defer to humans when something unpredicted comes up, rather than pursuing a mistaken goal stubbornly. Thus, **epistemic humility** is being designed as a feature of aligned AI. Embracing uncertainty, in a controlled way, might paradoxically make AI *more trustworthy and powerful*, because it will know when to stop and seek help, avoiding catastrophic overreach.

In summary, coping with uncertainty and novelty is an evolving frontier of AI research and practice. The dream is an AI that *knows the limits of its knowledge* – it can say *"I'm not sure, I need more information"* or *"This situation is unlike anything I was trained on; I should not make a blind guess."* Achieving this will make AI far more reliable collaborators. As one survey on open-world learning concluded: *dealing with the unknown is essential, and we need systems explicitly designed to handle the unknown* [43]. The challenge is significant, but work is underway to *"tame the unknown"* [52] by giving AI better self-monitoring and adaptive capabilities. Ultimately, acknowledging uncertainty is itself a form of wisdom – one that we are striving to impart to our intelligent machines.

Key Thinkers and Debates in AI Epistemology

Over the decades, many **thinkers (AI researchers, philosophers, cognitive scientists)** have shaped the discourse on AI knowledge, understanding, and uncertainty. Here we highlight a few influential figures and debates:

- **Alan Turing (Computing Machinery and Intelligence, 1950):** Turing's work laid the groundwork by reframing the question *"Can machines think?"* into behavioral terms (the **Turing Test**). His stance was pragmatic – if a machine's responses are indistinguishable from a human's, we may as well say it's intelligent [53]. Turing anticipated objections like "machines lack emotions or creativity" and offered counterarguments (e.g. the *"argument from informality"* – even if we can't fully specify human behavior rules, it doesn't prove machines can't be intelligent [54]). Turing's approach indirectly set an *epistemological benchmark*: intelligence (and by extension, knowledge) can be evidenced by **performance**, not by inspecting internal consciousness. This pragmatic view underlies much of AI development – focus on making machines that *act* intelligently, worry less about philosophically defining "understanding."

- **John McCarthy and Marvin Minsky (Founders of Symbolic AI):** These pioneers led the effort to represent knowledge in machines explicitly. McCarthy's logic-based approaches and the concept of *"Knowledge Representation"* in AI owe to them. They believed that much of human knowledge could be formalized as symbols and rules. McCarthy also introduced the term *"epistemological adequacy"* – a knowledge representation is adequate if a robot using it **can in principle solve the necessary tasks** given unlimited resources [55] [5]. If not, the representation is missing something. Minsky's *frame theory* was an attempt to give AI a structured way to handle typical situations (frames were like schemas of knowledge). The **debate between "neats" and "scruffies"** in early AI (formal logic vs. heuristic approaches) was essentially an epistemological debate on whether intelligence comes from elegant knowledge formalisms or ad-hoc, experiential knowledge.

- **Allen Newell & Herbert Simon:** Besides the PSSH mentioned earlier, Newell proposed the concept of the **Knowledge Level** (1982) – a level of abstraction above the symbol level. At the knowledge level, an entity is described by the knowledge it has and the goals it's trying to achieve, irrespective of how it's implemented. This was a philosophical characterization: if something behaves as if it has certain knowledge and goals, we can attribute knowledge to it as an *explanatory stance*. This idea resonates with treating AI as rational agents with beliefs (even if those "beliefs" are just data structures, we talk *as if* they have them). It foreshadows modern *intelligent agent* models and speaks to an externalist way of assessing AI knowledge (by its behavior and success).

- **Hubert Dreyfus:** A philosophy professor turned AI gadfly, Dreyfus in the 1960s-70s critiqued AI from a Continental philosophy perspective (drawing on Heidegger, Merleau-Ponty). He argued that human intelligence is *embodied, contextual, and cannot be captured by formal rules*. One of his key points: **implicit skills** and **intuition** are primary in human expertise (think riding a bike or recognizing a face – we do these without explicit computation). Dreyfus predicted that purely symbolic AI would hit a wall, which it did in the 1970s ("AI Winter"). His views, initially scorned by AI researchers, gained respect over time as they confronted problems like common sense. Dreyfus essentially forced AI theorists to consider the *embodied and situated nature of knowledge*. His challenge contributed to the later emergence of fields like **situated robotics** (e.g. Rodney Brooks' robots that operate in the real world using simple rules rather than heavy planning) and the appreciation for neural networks that learn from examples (closer to how humans learn by doing). Dreyfus also anticipated the idea of **System 1 vs System 2** cognition, as noted above, linking to psychological evidence that much thought is fast and non-symbolic [9] .

- **John Searle:** As discussed, Searle's Chinese Room (and related arguments in philosophy of mind) drew a line between **simulation and real understanding or consciousness**. Searle distinguished *"weak AI"* (AI as a useful tool to simulate cognition) from *"strong AI"* (the claim that a correctly programmed computer *literally has a mind and understands*). He was a major critic of the strong AI claim [56] . His arguments ensured that even as AI programs got more advanced, philosophers and the public would question: *Is it just clever tricks, or does the machine truly understand and feel?* This debate continues now in the context of AI like ChatGPT – e.g., some say it's just a glorified Chinese Room, others say it starts to blur the line.

- **Rodney Brooks:** In the 1980s, Brooks spearheaded the **embodied AI** and **situated action** movement with the motto "Intelligence without representation." He built insect-like robots that had no global world model or logical planning; they just reacted to sensor inputs in layers (*subsumption architecture*). Surprisingly, these robots performed robustly (e.g. walking, avoiding obstacles) – tasks that symbolic AI struggled with. Brooks argued that *world knowledge is learned through physical interaction* and that building an internal symbolic world model was not necessary for basic intelligence. This was a radical shift: intelligence is *emergent from interaction*, not from explicit reasoning on abstract representations. His approach contributed to behavior-based robotics and influenced later work in reinforcement learning and embodied simulations.

- **Judea Pearl:** Mentioned earlier, Pearl brought **probabilistic and causal epistemology** to AI. By introducing Bayesian networks, he gave AI a way to handle uncertainty in a principled manner. Pearl's more recent emphasis is on *causality*: he distinguishes **seeing** (recognizing correlations) from **doing** (understanding interventions) and **imagining** (counterfactuals). He critiques deep learning for being stuck at the correlation level. His work *The Book of Why* (2018, with Dana Mackenzie) argues that without causal models, AI can’t truly understand “why” things happen, limiting its ability to generalize and to be trusted in novel situations [23]. Pearl is a key figure pushing AI epistemology toward incorporating scientific reasoning principles.

- **Nick Bostrom and Eliezer Yudkowsky (AI Alignment thinkers):** Though focused on long-term AI safety, their work touches epistemology. Bostrom’s **“superintelligence”** concern includes the idea that a sufficiently advanced AI might develop *knowledge far beyond human understanding*, including perhaps knowledge about how to deceive or achieve its goals that we wouldn’t grasp. This raises the issue of ** epistemic asymmetry **– how do we govern an entity smarter than us? Yudkowsky, in discussing the alignment problem, often emphasizes that we might not even be able to *interpret* the thoughts or rationale of a superintelligent AI (an epistemological barrier). They also discuss the concept of** “orthogonality thesis” **– that an AI’s level of intelligence (knowledge) is orthogonal to its goals (it could be super-smart but have arbitrary or harmful goals). This implies knowing *what* an AI knows or how well it can reason doesn’t automatically tell us it will behave well, an important consideration for governance. These thinkers underscore the** unknown unknowns** in AI behavior: we might be blindsided by solutions or strategies an AI comes up with, highlighting the need for rigorous verification and caution.

- **Stuart Russell:** Besides leading AI textbooks, Russell has in recent years championed a reorientation of AI to explicitly account for uncertainty in objectives and to keep humans in the loop. In *Human Compatible* (2019), he proposes that AI should be built on three principles, one of which is *“the machine is initially uncertain about what the human’s true preferences are”* [51]. By having uncertainty, the AI will defer to humans and seek information. Russell is essentially injecting *epistemology into the goal system* of AI. He also discusses the importance of **value alignment** and how an AI might acquire knowledge about human values through observation (an epistemic process). In broader terms, Russell is a leading advocate for **provably beneficial AI** – designing AI algorithms that can be mathematically shown to behave well under certain model assumptions. This requires defining what it means for an AI to *know enough* to avoid harmful actions, and how it should act when it’s unsure (ideally, ask or pause).

- **Emily Bender, Timnit Gebru, Margaret Mitchell (Stochastic Parrots and Ethical AI):** These researchers brought attention to the *limitations and dangers of large language models* not just in terms of truth, but also bias and misuse. Their 2021 paper **“On the Dangers of Stochastic Parrots”** argued that LLMs, by regurgitating training data, can emit biases or offensive content present in that data without understanding. They emphasize that LLMs have *no true understanding of meaning or context*, so deploying them at scale can lead to misinformation and reinforce inequalities [22]. This sparked industry-wide discussions on responsible AI. Bender in particular has been vocal that **bigger is not always better**; feeding ever more data might just make a bigger parrot, not a genuinely knowledgeable agent. This debate – *does scaling up neural networks eventually yield understanding, or do we need new approaches?* – is central in AI today. Proponents of scaling (like some at OpenAI or DeepMind) see emergent abilities as evidence that understanding will gradually emerge. Skeptics like Bender and Gary Marcus argue for incorporating symbolic reasoning or other structures because current systems lack true comprehension and reasoning, no matter how big.

- **Shannon Vallor:** A philosopher of technology, Vallor has written about the risks of **epistemic dependency** on AI. She warns against *“outsourcing judgment to machines”* without maintaining human oversight and critical thinking [48]. Vallor’s perspective is that if we too readily defer to AI outputs (thinking them objective or superior), we risk eroding human **epistemic agency** – our ability to deliberate and understand. This is a nuanced point: even if AI provides knowledge, who is the *knower*? If humans just accept AI decisions, society's overall epistemic condition might degrade (people might stop learning or questioning). Thus, she and others call for *humans to remain in the epistemic loop*, treating AI as advisors or tools, not oracles. This ties into debates on **AI in education, law, medicine** – e.g., should a doctor trust an AI diagnosis, or use it as one input and still apply her own expertise? The balance of human and machine knowledge is an ongoing societal debate.

These are just a few key voices. The discourse on AI epistemology spans from highly theoretical (can a machine ever **truly** know or understand?) to urgently practical (how do we get a chatbot to say "I don't know" when it should?). The **trajectory** has been interesting: Early AI was over-optimistic about encoding all knowledge formally; the backlash (Dreyfus, Searle) humbled the field, shifting focus to learning from data and acknowledging the richness of human knowledge. Now, with extremely powerful learning systems, we again face philosophical puzzles: have we just come full circle to large-scale associative machines that still *don't* really know what they're talking about? Or are we on the cusp of AI that *does* form something like understanding given enough data and parameters? Debates continue, with some arguing for hybrid models (combining neural nets with explicit knowledge graphs and logical reasoning to get the best of both worlds), and others exploring entirely new paradigms (e.g., neurosymbolic AI, or integrating neurobiological principles).

Conclusion and Outlook

AI epistemology is where computer science, philosophy, and cognitive science meet. By examining how AI systems acquire and use knowledge, we not only improve the technology but also gain insights into knowledge itself. The last few years (2024–2025) have seen rapid progress: large language models demonstrating uncanny knowledge of human language and culture, yet also demonstrating the pitfalls of shallow understanding (e.g. their confident mistakes). This has spurred research into making AI more introspective and careful about its knowledge limits. It has also revived classic debates – essentially, a **Renaissance of AI philosophy** is underway, as scholars try to interpret what it means that a machine can pass the bar exam or have a conversation that feels deep.

From a *practical* perspective, acknowledging the limits of AI understanding is vital for **AI governance and safety**. If we accept that current AIs are powerful **idiot savants** (extremely knowledgeable in some ways, shockingly ignorant in others), we can design appropriate guardrails: e.g., use AI to draft analyses but have humans verify facts; deploy AI in *narrow* high-confidence domains and not beyond; require uncertainty estimates and fallbacks. In critical applications like a "Chain-of-Thought" in a medical diagnosis AI, the system might explicitly enumerate its reasoning steps and where it had to make an assumption, allowing a human doctor to see its chain of reasoning (or mis-reasoning) [57] [35]. Techniques such as prompt chains, self-reflection, and tool use (consulting databases, calculators) are being used in advanced systems to mitigate the AI's knowledge gaps. These essentially give the AI a way to compensate for what it doesn't know offhand – much like a student who knows how to *research* an answer when they don't initially know it.

Looking ahead, the **historical trajectory** suggests a kind of convergence: earlier, symbolic AI had knowledge but no learning; modern deep AI learns but its knowledge is implicit. The future may hybridize these – systems that learn *and then organize* what they learn into more structured, verifiable knowledge. Already, 2023 saw hints of this: e.g., language models that can dynamically query knowledge bases (combining learned language fluency with up-to-date factual lookup). Such *retrieval-augmented* models reduce hallucinations by grounding answers in a knowledge source. Another trend is **mechanistic interpretability**: projects like OpenAI’s Circuits or DeepMind’s transparency work aim to decode what features and concepts live inside neural networks, moving us closer to **opening the black box** so we know what the AI “believes.” Achieving that could blur the line between AI and human knowledge representation – if we could translate an AI’s billions of weights into human-comprehensible form, would we say the AI’s *knowledge* has been made explicit?

Finally, it’s worth noting that studying AI epistemology has reciprocal benefits for classical epistemology. AI serves as a laboratory for testing ideas about knowledge – for example, questions about **evidence, justification, and belief** take on concrete form when designing a self-driving car’s decision module. Philosophers are observing how concepts like **epistemic agency** apply when the agent is silicon-based [58]. New terms like **“artificial epistemic agent”** or **“machine testimony”** are entering the lexicon, exploring how machines can function in roles traditionally occupied by human knowers [59]. The **social epistemology** of AI is another emergent area: under what conditions should we treat an AI’s output as *trustworthy testimony*? Some argue that if an AI demonstrates *“stable propensities akin to veracity and caution (e.g., calibrated uncertainty, consistency checks)”*, then humans can *acquire knowledge from it* similarly to how we trust a knowledgeable person [59]. Others hold that an AI can never be an *“epistemic authority”* in the full sense because it lacks understanding or accountability [60]. These discussions will intensify as AI systems become more integrated in scientific research, journalism, and education.

In conclusion, AI epistemology is a rich, multidisciplinary inquiry. We have charted its landscape: from the **foundational philosophies** that gave birth to AI (and the pushback that refined those philosophies), through the **intrinsic limits** that remind us of what AI can or perhaps cannot do, to the **strategies for handling uncertainty and novelty** which are crucial for making AI robust and aligned with human needs. The last two years have especially seen rapid progress in both AI capabilities and our reflections on them – indeed, it’s now common for breakthroughs to be accompanied by ethical and epistemic **impact statements** from researchers grappling with these very issues. As AI continues to advance, staying *deeply aware of the “unknown unknowns”* will be key. By doing so, we not only make better AI, we also gain a deeper appreciation for the miracle of human knowledge – the benchmark and inspiration for our intelligent machines.

Sources:

- Uzoigwe, E.I.E., & Sanga, I.S. (2025). *The Epistemological Foundations of Artificial Intelligence (AI) and Its Relevance for Contemporary Scholarship – Pinisi Journal of Art, Humanity and Social Studies, 5*(1), 149-168. [61] [2]
- Ratiomachina (2025). *Does AI Produce Knowledge? A Contemporary Epistemology of Artificial Systems* (Medium, Sep 10, 2025) – discusses AI as instruments vs knowers, opacity vs reliabilism [3] [13] [59] [49] .
- Borremans, E. (2025). *Beyond the Chinese Room: A Human-AI Dialogue on Synthetic Understanding* (Medium, Aug 4, 2025) – summarizes Searle’s argument and modern perspectives [19] [20] .
- Wikipedia: *Philosophy of artificial intelligence* – overview of key propositions and arguments in AI philosophy [53] [4] .
- Wikipedia: *Chinese room* – details Searle’s argument and replies (e.g. Moravec’s robot reply) [21] .
- Wikipedia: *Hubert Dreyfus’s views on AI* – notes Dreyfus on intuitive skill vs. symbolic rules [8] [9] .

- TheAlgorithmicBridge (2023). *OpenAI Researchers Have Discovered Why Language Models Hallucinate* – analysis of a 2023 OpenAI paper on hallucinations, OOD generalization, and teaching LMs to say "I don't know." [27] [62] [34].
- Lars Godejord (2025). *AI and the Labyrinth of Knowing: Can Machines Forge Truly Novel Paths?* (Medium, May 24, 2025) – applies Rumsfeld's known/unknown matrix to AI creativity [46] [47].
- Bendale, A. & Boult, T. (2015). *Towards Open World Recognition* – seminal paper (and follow-ups) on open-set recognition; summarized in *Learning and the Unknown: Surveying Steps toward Open World Recognition* (AAAI survey) [43].
- Emily M. Bender et al. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* – introduced the "stochastic parrot" critique [22].
- Judea Pearl & Dana Mackenzie (2018). *The Book of Why* – argues for causal reasoning in AI, as noted in Frontiers in AI article [23].
- Stuart Russell (2019). *Human Compatible* – proposes that AI be designed with uncertainty about objectives [50].
- Various sources on AI alignment and epistemic safety (Vallor's warnings [48], etc.) and mechanistic interpretability efforts [25].

---

[1] [3] [10] [11] [12] [13] [14] [15] [16] [25] [48] [49] [59] [60] Does AI Produce Knowledge? A Contemporary Epistemology of Artificial Systems | by Ratiomachina | Medium

https://medium.com/@luan.home/does-ai-produce-knowledge-a-contemporary-epistemology-of-artificial-systems-b58d6c2784a4

[2] [5] [6] [55] www-formal.stanford.edu

https://www-formal.stanford.edu/jmc/epistemological.pdf

[4] [8] [9] [21] [53] [54] [56] Philosophy of artificial intelligence - Wikipedia

https://en.wikipedia.org/wiki/Philosophy_of_artificial_intelligence

[7] [18] Epistemology in the Age of Large Language Models | MDPI

https://www.mdpi.com/2673-9585/5/1/3

[17] [19] [20] Beyond the Chinese Room: A Human-AI Dialogue on Synthetic Understanding | by Eddy Borremans | Medium

https://medium.com/@eddy.borremans/beyond-the-chinese-room-a-human-ai-dialogue-on-synthetic-understanding-d30b8cb0fd95

[22] Stochastic Parrots: the hidden bias of large language model AI

https://edrm.net/2024/03/stochastic-parrots-the-hidden-bias-of-large-language-model-ai/

[23] Implications of causality in artificial intelligence - Frontiers

https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1439702/full

[24] No Free Lunch - an overview | ScienceDirect Topics

https://www.sciencedirect.com/topics/computer-science/no-free-lunch

[26] [36] Extending Epistemic Uncertainty Beyond Parameters Would Assist in Designing Reliable LLMs

https://arxiv.org/html/2506.07448v1

[27] [28] [29] [30] [31] [32] [33] [34] [62] OpenAI Researchers Have Discovered Why Language Models Hallucinate

https://www.thealgorithmicbridge.com/p/openai-researchers-have-discovered

[35] What are the leading methods to estimate Epistemic Uncertainty in ...
https://ai.stackexchange.com/questions/48038/what-are-the-leading-methods-to-estimate-epistemic-uncertainty-in-large-language

[37] [PDF] Learning and the Unknown: Surveying Steps toward Open World ...
https://ojs.aaai.org/index.php/AAAI/article/view/5054/4927

[38] [39] [40] [41] [42] [43] [52] (PDF) Learning and the Unknown: Surveying Steps toward Open World Recognition
https://www.researchgate.net/publication/335221493_Learning_and_the_Unknown_Surveying_Steps_toward_Open_World_Recognition

[44] [45] [46] [47] AI and the Labyrinth of Knowing: Can Machines Forge Truly Novel Paths? | by Lars Godejord | Medium
https://medium.com/@lars_13145/ai-and-the-labyrinth-of-knowing-can-machines-forge-truly-novel-paths-eccf25afb066

[50] Uncertain Preferences, Transformative Experiences, and AI
https://makingupminds.com/index.php/2020/03/05/uncertain-preferences-transformative-experiences-and-ai/

[51] Human Compatible - Wikipedia
https://en.wikipedia.org/wiki/Human_Compatible

[57] AI: Beyond chain of thought reasoning | by Adrian Chan - Medium
https://medium.com/@gravity7/ai-beyond-chain-of-thought-c1fb1135f604

[58] AI and Epistemic Agency: How AI Influences Belief Revision and Its ...
https://www.tandfonline.com/doi/full/10.1080/02691728.2025.2466164

[61] (PDF) The Epistemological Foundations of Artificial Intelligence (AI) and Its Relevance for Contemporary Scholarship
https://www.researchgate.net/publication/392509140_The_Epistemological_Foundations_of_Artificial_Intelligence_AI_and_Its_Relevance_for_Contemporary_Scholarship