TOWARDS ETHICAL AI

Hugo Latapie
hugo@taijituai.com

ABSTRACT

Summary. Artificial intelligence (AI) has seen rapid adoption across diverse applications, prompting discussions on "Ethical AI." This paper explores two key dimensions: (1) whether the *behavior* of current AI systems can be labeled "ethical" or "unethical," given their lack of common sense; and (2) how AI systems are *represented* to end users to ensure they do not overtrust or misuse these tools. Drawing from a "competent to stand trial" analogy in law, we conclude that present-day AI, although powerful for many tasks, is insufficiently context-aware to bear moral agency. Thus, ethical considerations must focus on labeling and disclaimers that clarify the AI's limitations and the need for human oversight, particularly in mission-critical scenarios.

***K*eywords** Artificial Intelligence · Common Sense · Ethics · Competent to Stand Trial · AI Representation

1 Introduction

Artificial Intelligence (AI) has undergone a rapid evolution, reflected in a wide array of real-world deployments. From recommendation engines to semi-autonomous driving, advancements have prompted greater attention to questions of "Ethical AI." Yet the term "ethical" can carry multiple interpretations. On one hand, there is the *behavior* of AI systems themselves: Do these outputs reflect moral reasoning or can they be unethical? On the other hand, there is how these systems are *represented* to users: Are users given accurate information about the AI's capabilities so they do not overtrust or misuse the system?

We draw upon an analogy to "competence to stand trial" [Turing, 1950, Floridi and Sanders, 2004]—in legal terms, a defendant must understand the charges, communicate with counsel, and make rational decisions to be held accountable. Likewise, an AI lacking contextual understanding (common sense) may appear to make decisions, but it is not truly "competent" in the ethical sense. Instead, the obligation for ethical deployment lies with those who build and label the AI. In prior work, "Common Sense Is All You Need" [Latapie, 2025a], we have defended emphasizing minimal prior knowledge, flexible learning, and adaptive reasoning as prerequisites for robust autonomy.

2 Defining AI Competence: Common Sense as the Missing Piece

2.1 Common Sense in AI: Why Contextual Understanding Matters

Common sense refers to an AI's ability to interpret context, adapt to novel conditions, and anticipate outcomes, aligning with broader definitions of intelligence [Legg and Hutter, 2007, Russell and Norvig, 2010]. For humans, it arises naturally. For most AI, progress primarily involves data-driven modeling in constrained domains. Lacking context, these AI solutions can produce strikingly accurate results in typical cases but fail dramatically when faced with edge cases or moral ambiguities.

2.2 Consequences of Incompetence: AI Behaviors vs. Human-Like Agency

When an AI lacks the robust contextual reasoning that humans possess, its "decisions" may be powerful heuristics but do not arise from moral deliberation [Allen et al., 2000]. Analogous to a defendant who cannot comprehend the charges or develop a defense, we cannot meaningfully label uncomprehending AI outputs as "ethical" or "unethical."

2.3 Referencing Our Prior Work on Common Sense

We have previously emphasized that minimal prior knowledge and flexible learning are keys to building truly autonomous AI [Latapie, 2025a] [Latapie, 2025b]. Extending that logic, an AI must integrate contextual reasoning and consequence awareness to bear moral accountability for its behavior.

3 Distinguishing AI System Behavior from Ethical Agency

3.1 The Legal Analogy: Incompetent to Stand Trial

Table 1: Core parallels between legal competence and AI's need for common sense.

Legal Competence	AI Common Sense Requirements
Understanding (charges/process)	Contextual Learning
Communication (with counsel)	Interaction with Users/Data
Rational Decision-Making	Adaptive Reasoning and Consequence Awareness

A defendant deemed incompetent to stand trial cannot be morally or legally accountable to the usual degree [Turing, 1950, Floridi and Sanders, 2004]. Likewise, an AI system lacking context should not be expected to make "ethical" decisions, as it does not grasp ramifications or moral complexity.

3.2 A System Without Common Sense: Behavior vs. Intent

Chatbots generating plausible yet incorrect statements (sometimes called "hallucinations" [Shaikh, 2024]) are not intentionally deceptive in the moral sense. They are pattern matchers, lacking the contextual knowledge needed to understand truth-value. Accusations of "lying AI" conflate text generation with genuine dishonesty.

3.3 Illustrative Scenarios: Confusions Around Levels of Autonomy

Consider a semi-autonomous driving system labeled "Full Self-Driving" (SAE Level 2), which demands driver responsibility. Many customers, becoming comfortable under normal conditions, incorrectly treat it like a Level 4 system with remote support. This misattribution of competence can lead to dangerous overreliance, highlighting that the system itself is not morally at fault; it is incompetent from an ethical standpoint, leaving accountability to humans.

4 Representing AI Systems to Customers

4.1 The Significance of Labeling and Transparency

AI systems without contextual reasoning can be highly valuable, e.g., producing creative ideas, analyzing large datasets, or assisting with routine tasks. The risk is overestimating their reliability or ethical grounding. Candid disclaimers about potential "hallucinations," data biases, and operational domains help ensure end users do not ascribe moral agency to the AI.

4.2 "Ethical AI" in Marketing vs. Reality: Pitfalls of Overpromising

Organizations might highlight "ethical guidelines" while marketing AI solutions in ways that oversell capabilities. Labeling an AI system as "responsible" or "fair" can mislead customers into thinking it has moral awareness. However, as it lacks common sense, failures in novel conditions could produce unethical outcomes [European Commission, 2019, IEEE, 2019].

4.3 Case Examples: Fun, Creative, or Risky "Hallucinations"

A comedic chatbot that invents nonsensical content can be entertaining if a user expects random creativity. Problems arise if that same user relies on it for factual advice. Clear labeling states that such "fabrications" may occur, preventing unscrutinized trust in the AI's behavior.

5 The Risks of Misrepresentation

5.1 Overtrust: When Customers Believe AI Is More Capable Than It Is

When marketing or hype suggests the AI is "truly ethical," or "trustworthy," organizations risk fostering overconfidence on the part of users. This can be especially perilous in real-time decision contexts, such as factory automation, financial risk modeling, or personalized medicine, where system errors have serious repercussions.

5.2 Real-World Consequences: Mission-Critical Tasks with an "Incompetent" AI

- **Industrial Control:** An AI system that lacks common sense as defined in Latapie [2025b] might misread sensor anomalies, leading to catastrophic shutdowns or insufficient safety measures.
- **Financial Automation:** Automated loans or credit scoring used beyond tested demographics can propagate discrimination or undue harm.

5.3 Mitigation Strategies: Clear Boundaries, Human Oversight, and Safety Mechanisms

1. **Clear Boundaries:** Publicly describe exact contexts in which the AI is considered reliable and provide clear metrics in easy-to-understand language like food labels.
2. **Human Oversight:** Skilled professionals remain the final checkpoint for decisions that might have reputations, business, moral or safety weight.
3. **Fallback Procedures:** Provide immediate override systems or "kill switches" to avert calamities when the AI proves incompetent.

6 Toward Ethical AI: Merging Competence with Honest Communication

6.1 Formalizing AI Competence: Steps Toward Common Sense Integration

Ongoing research in multi-modal learning, reinforcement learning, and symbolic integration aims to give AI a deeper sense of context [Chollet, 2019, Russell and Norvig, 2010]. Achieving robust common sense would allow AI to handle unexpected events and weigh moral factors, forming the backbone of an "ethical" agent.

In our previous works [Latapie, 2025a,b], we argued for a minimal prior knowledge approach—coupled with contextual learning and adaptive reasoning—to systematically build AI competence. These principles outline how an AI can develop real-world awareness rather than merely pattern-matching within constrained domains, thereby reducing hallucinations and misinterpretations. By integrating such common sense frameworks, AI systems move closer to satisfying the foundational requirements for moral agency or "ethical" action.

6.2 Aligning System Behavior with Clear Representation

If an AI's competence is limited, labeling it accurately avoids luring users into false assumptions. For instance, disclaiming that "this system may produce creative fictions" or requiring user supervision in auto-driving fosters safer usage.

6.3 Anticipating Future Developments: As AI Gains Competence, Ethical Agency Becomes Relevant

As computing power and AI architectures continue to scale, systems are positioned to handle increasingly complex tasks and domains. This heightened ability raises the stakes for how these systems reason about context, consequences, and moral factors. Without incorporating robust common sense, even advanced AI will risk deploying superficial heuristics that lack ethical grounding.

Only after surpassing a threshold of contextual awareness might AI be expected to "act ethically" in a moral sense. At present, "ethical AI" must be construed primarily as *human* ethical diligence in disclaiming, limiting, and monitoring incompetent AI to prevent harm. In the future, should large-scale AI models integrate core principles of common sense and adaptive reasoning, they may begin to bear some responsibility for the outcomes of their own decisions, thereby evolving toward genuine ethical agency.

7 Conclusion and Outlook

7.1 Recap: Behavior vs. Representation

Our discussion "Towards Ethical AI" highlights two main themes:

1. The nature of an AI system's outputs, which cannot be deemed morally or immorally guided in the absence of contextual understanding.
2. The manner in which AI capabilities are *represented* to customers, often conflated with genuine agency and ethics.

7.2 Reinforcing the "Competent to Stand Trial" Analogy

As an incompetent defendant remains unaccountable in normal legal standards, an AI lacking common sense should not shoulder "ethical" or "unethical" labels. We do not condemn the AI as malicious—rather, the duty lies with developers, marketers, and regulators to position, label, and oversee these systems responsibly.

7.3 Call to Action: Practical Measures for Honest Positioning and the Road Ahead for Common Sense AI

- **Honest Labeling:** Explicitly outline domain boundaries, disclaim potential fabrications, and confirm it is not a moral agent.
- **Appropriate Deployment:** Avoid handing high-stakes decisions to an AI that cannot interpret the deeper context of those choices.
- **Further Research:** Push for advancements in common sense reasoning to inch closer to an AI that might eventually earn the title "ethical agent."

By bridging the gap between actual AI competence and user perception, organizations can reap the benefits of today's AI without glossing over limitations. Over time, continuing research in contextual reasoning may enable AI to pass a bar akin to legal competence, rendering ethical AI more than a marketing slogan.

References

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

Luciano Floridi and J. W. Sanders. On the morality of artificial agents. *Minds and Machines*, 14(3):349–379, 2004.

Hugo Latapie. Common sense is all you need. arXiv preprint, 2025a. arXiv:2501.06642.

Shane Legg and Marcus Hutter. A collection of definitions of intelligence. arXiv preprint, 2007. arXiv:0706.3639.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, Upper Saddle River, NJ, 3rd edition, 2010.

Colin Allen, Wendell Wallach, and Iva Smit. Why machine ethics? *IEEE Intelligent Systems*, 21(4):12–17, 2000.

Hugo Latapie. Further reflections on common sense for ai. arXiv preprint, 2025b. arXiv:2502.12345.

Kaif Shaikh. AI can strategically lie: From innocent errors to lying, manipulation, and deception. `https://interestingengineering.com/culture/truth-about-ai-deception`, 2024. This article sheds light on the growing capabilities of AI to deceive, strategize, and mislead, highlighting the urgent need for stricter ethical controls.

European Commission. Ethics guidelines for trustworthy AI. `https://ec.europa.eu/futurium/en/ai-alliance-consultation`, 2019. Accessed May 2023.

IEEE. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. `https://ethicsinaction.ieee.org/`, 2019.

Francois Chollet. On the measure of intelligence. arXiv preprint, 2019. arXiv:1911.01547.