Toward a Dispassionate Philosophy of AI

International Online Session

"Human Society and AI – How is our Co-Existence possible based on irrefutable Truth and general Good?"

February 27, 2026

Romaric Jannel

Ritsumeikan University

Collège international de philosophie

The Current Landscape

- Since ChatGPT's release, philosophy of AI has polarized into two tendencies:
  - Risk-focused: Emphasizes real, potential, and imagined risks AI poses to humanity
  - Optimistic: Incorporates entrepreneurship and corporate ethics, depicting AI as revolutionary for the greater good

Synthese (2025) 206:277
https://doi.org/10.1007/s11229-025-05378-9

ORIGINAL RESEARCH

The negativity crisis of AI ethics

Peter Königs[1]

Received: 13 May 2025 / Accepted: 11 November 2025

Abstract
Despite the great positive potential of AI, the AI ethics community has presented a rather gloomy picture of AI's ethical implications. This paper examines the negativity within AI ethics through a philosophy of science lens. The prevailing negativity is a result of the particular way the discipline is institutionally organized, which pressures AI ethicists to portray AI in a critical light. As a consequence, the overall picture of AI offered by the AI ethics community is one-sided and negatively biased. We should be skeptical about the negative narrative promoted by AI ethics and explore ways of reforming the system.

A Balanced Position

- Murray Shanahan (Emeritus Professor of AI at Imperial College London) suggests the truth lies between these two extremes.
- A more balanced position:
  - Dispassionate Philosophy of AI
  - Not indifferent or politically neutral, but examining our own inclinations before debating the technology

The Three Poisons Framework

Buddhist framework of three basic mind-states causing suffering:

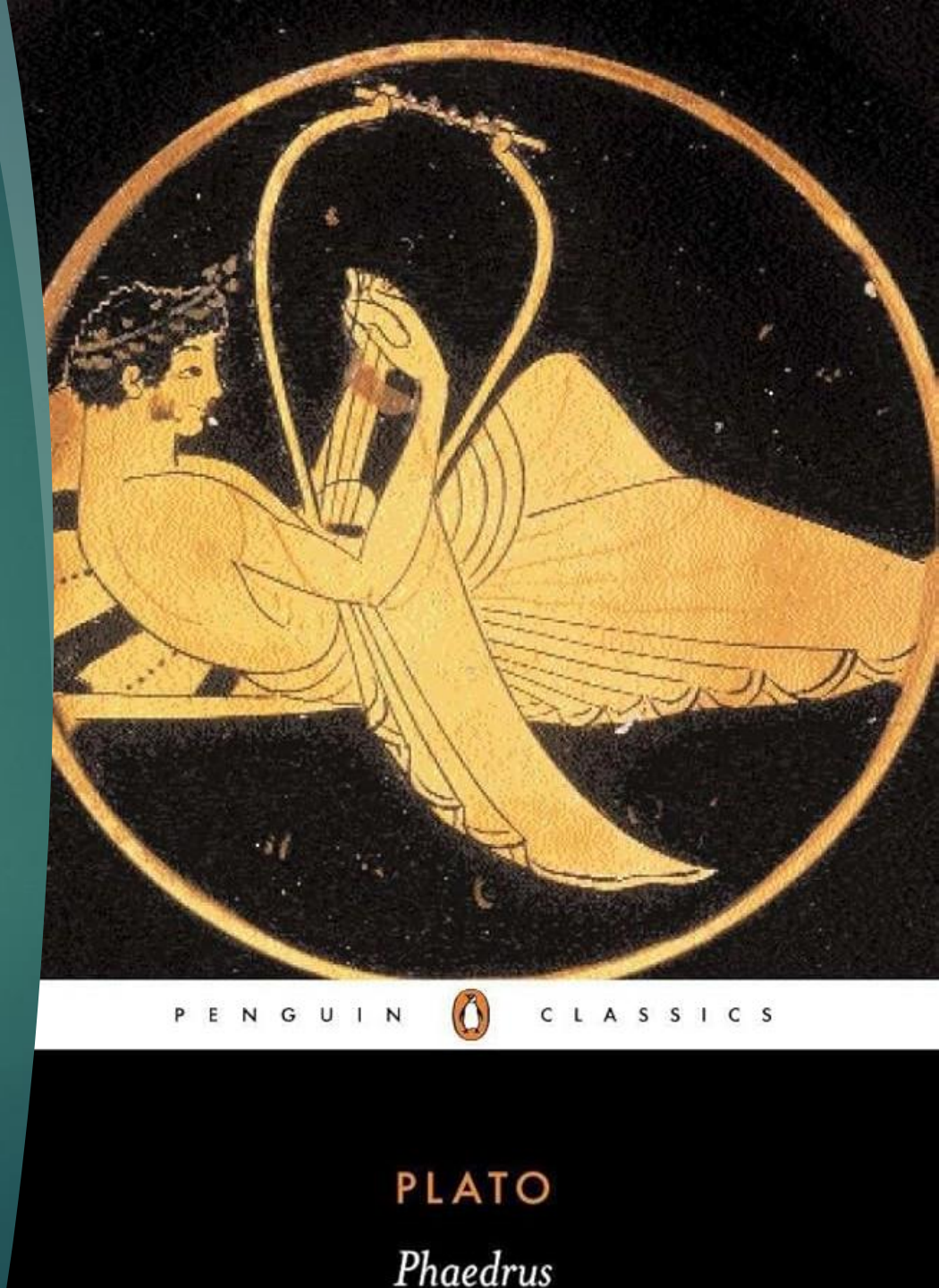
- Ignorance: Lack of clarity about what AI systems are and do
- Greed: Extracting value—profit, productivity, or prestige—faster than understanding consequences
- Aversion: Moral panic treating AI as a homogeneous object of fear
- A dispassionate philosophy begins by recognizing these three patterns in ourselves and our institutions, as they often influence the debate more than arguments do

Historical Analogy: Plato's King Thamus

Plato's Egyptian myth warns about new technologies:

- Writing will weaken memory
- Produce an appearance of wisdom without understanding
- People repeat marks without comprehension

- Writing became foundational for science, law, and coordination
- When knowledge is externalized, surrounding practices must adapt or the technology may reshape them in unwanted ways



AI: The Next Chapter

One critical difference from writing:

- Writing stores and transmits information
- AI systems produce outputs appearing as explanations and expertise
- Risk: societies lose stable signals of epistemic trust

Societies lose clarity on:

- Who actually knows?
- Who verified?
- Who is accountable?

Conclusion

A dispassionate philosophy of AI:

- Pays closer attention to potential and risks
- Treats polarization as symptoms of three issues
  - Ignorance about system abilities and limits
  - Greed in deployment strategies
  - Aversion as a psychological defense

The Path Forward

Following King Thamus as our guide:

- Not to stop disruption (societies learn through it)
- Prevent disruption from becoming destructive
- Preserve epistemic responsibility and trust
- Maintain conditions for accountability

AI and Ethics (2026) 6:13
https://doi.org/10.1007/s43681-025-00839-w

RESEARCH

Trustability and trustworthiness: conceptual foundations and the case of AI

Romaric Jannel[1] · Jonathan Tallant[2]

Received: 22 September 2025 / Accepted: 12 November 2025

Abstract

This paper distinguishes between trustability and trustworthiness as two conceptually and normatively distinct conditions for legitimate trust, a distinction that has been largely absent from the philosophical literature. Trustworthiness concerns whether an agent merits trust, while trustability names a prior condition: whether the entity in question is even the kind of thing to which trust can coherently apply. Focusing on Faulkner's grammar of trust and recent work by Massaguer Gómez on human–robot interaction, we argue that many artificial intelligence systems today elicit trust without being trustable—a category error with ethical consequences for the design, deployment, and governance of emerging technologies. We propose that trustability functions as a normative threshold for evaluating whether trust in AI is not only misplaced but also structurally incoherent, and we show how this concept allows us to differentiate between merely instrumental reliance and genuinely normatively structured trust. This clarifies when evaluations should shift from "trust" to reliance with accountability. We also examine relevant philosophical discussions that have anticipated parts of our argument, situating our approach within debates about trust in governments, institutions, and nonhuman agents, and clarifying how our framework builds upon and departs from existing positions. Finally, we explore the conceptual and institutional conditions under which future AI systems might become both trustable and trustworthy—outlining technical, moral, and political prerequisites for such a development. This distinction provides a framework for diagnosing inappropriate trust, clarifying when it is possible, when it is normatively justified, and when it is conceptually impossible.

Keywords Trust · Trustworthiness · Trustability · Reliance · Institutional surrogate trust · Governance · Artificial intelligence

Thank You!