

TRUSTWORTHY ARTIFICIAL INTELLIGENCE: the relevance of interdisciplinary approach.

In the first quarter of the 21st century, the capacity to produce, process and distribute data has grown exponentially, driven by computing and digital communication technologies. In this context, artificial intelligence (AI), a set of disciplines born alongside electronic computing in the mid-20th century, has achieved a remarkable development. Although AI is not a new idea, its current impact following the application of Generative AIs (GenAI) is broader than ever, reaching various areas of society and culture. The advance of this type of applied artificial intelligence, mainly Pre-Trained Large Language Models (PT-LLM) on Artificial Neural Networks (ANN) under Machine Learning/Deep Learning (ML/DL) techniques, represents great opportunities for improvement in various fields of work and study. But it also entails significant challenges for the redesign of roles and jobs and the development of capabilities to allow a harmonious incorporation of these tools with the least possible liabilities in the change management.

In its quest for knowledge, science combines processes of discovery with research methods that produce theoretical models in various formats. The application of this knowledge as a transformative tool has the potential to improve human life in various areas, but also to generate collateral damage, environmental impact and complex social dilemmas. To reflect on the *trustworthiness* of a particular discipline involves questioning not only about the context and means for obtaining its knowledge, but also about its limits and the responsibilities that emerge from its development and application. In this regard, the epistemology of science serves as a guiding framework that seeks to ensure that research and its applications are aligned with fundamental concepts and proper procedures.

In line with this, Karl Popper argues that all science begins with philosophical problems and ends with philosophical problems. AI does not seem to be an exception. A cursory glance at the seminal writings of the founders and the developments of pioneers in the field reveals that their original concerns are of a philosophical and epistemological order. In a second stage, the formulation of derived problems in conjunction with the search for scientific answers establishes, more or less organically, some substantive theories of AIs. This somewhat fragmentary theoretical framework –and still lacking in general consensus– constitutes, together with the evolving operational theories, the technical scaffolding that supports AIs as formal objects and technological artifacts.

In fast evolving disciplines such as AI, where the technical challenges are profound and the frontiers of knowledge are yet to be demarcated, there is a tendency to at least partially identify theoretical and epistemological discussions such as the concern for “trust” with the discussion of the ethical implications of the field. Contrary to this tendency, the aim should be to delve into the *quid est* of AI as the first and necessary question preceding any axiological assessment. This rapid “flight to ethics” often leaves aside the underlying scientific problems and shifts the debate towards “more accessible” practical issues. But focusing on some ethical dilemmas without establishing a solid theoretical framework for trust –one capable of making explicit the technical challenges that persist as unresolved obstacles– results in an exchange of “discourses” that is of little benefit to trustworthy artificial intelligence. AI trust and AI ethics are close related but they are not convertible.

For example, Nick Bostrom argues that the development of artificial general intelligence (AGI) or systems that truly emulate human intelligence remains an outstanding challenge and numerous “ethical” debates are actually preempting as yet unresolved technical issues. Despite advances in deep learning (DL), artificial neural networks (ANN) and natural language processing (NLP), the problems of intelligence, intentionality and common sense remain slow progressing and still opaque areas of research. Faced with this difficulty, some theorists tend to move towards the ethical aspect of AI, where discussions can be more speculative and do not require rigorous empirical validation.

Additionally, Stuart Russell points out that while the study of the ethics of AI is valuable, it can sometimes overshadow more pressing issues related to the technical control of systems. Rather than addressing concrete challenges, such as optimizing algorithms or solving alignment problems, ethical reflection allows theorists to explore abstract dilemmas without any commitment to tangible results. In this regard, debates about “the rights of robots” or “artificial consciousness” represent an intellectual escape route from the inherent complexity of improving systems that still face difficulties in basic reasoning and planning tasks.

Luciano Floridi argues that much of the ethics of AI is based on anthropomorphic analogies that, although appealing, do not address the real technical problems of the field. AI popularizers, influenced by popular culture narratives, tend to project human qualities onto AI, attributing intentions and values to it. This tendency to anthropomorphize results in AI trust debates regarding hypothetical scenarios –such as a machine takeover– that seem more appealing than the analysis of the inherent limitations of current models of representation or reinforcement learning.

Furthermore, many ethical debates lack the necessary technical rigor and focus on futuristic scenarios instead of addressing immediate challenges. Influenced by popular culture and science fiction narratives, the discourse about AI often leans towards speculative dimensions, pushing real structural problems into the background. Media attention is more easily captured by the discussion of ethical issues than by strictly technical debates. The development of a trustworthy artificial intelligence is fundamental to guarantee a humanistic vision of AI and to favor responsible innovation. It should stem from a primary consideration: that all available efforts and technical means be used to ensure that AIs meet criteria of governability, explainability, traceability and transparency.

In this context, it is equally necessary to consider interdisciplinary integration as a cornerstone for aligning technical developments with ethical, legal, and societal frameworks. The development of trustworthy artificial intelligence cannot be grounded solely in engineering or computational logics; it must also engage with philosophical, legal, psychological, and cultural dimensions. Humanistic knowledge—rooted in critical thinking, historical awareness, and ethical reflection—offers the conceptual tools needed to question assumptions, interpret social impacts, and foreground the dignity of the human person. When combined with technical expertise, it enables a more holistic and anticipatory approach to AI design and deployment, one that respects complexity, reduces unintended harm, and fosters public trust.