Navigating the Ethics of Artificial Intelligence

Subjects: Ethics

Contributor: Jack Harris , Veljko Dubljević

This entry delineates artificial intelligence (AI) ethics and the field's core ethical challenges, surveys the principal normative frameworks in the literature, and offers a historical analysis that traces and explains the shift from ethical monism to ethical pluralism. In particular, it (i) situates the field within the trajectory of AI's technical development, (ii) organizes the field's rationale around challenges regarding alignment, opacity, human oversight, bias and noise, accountability, and questions of agency and patiency, and (iii) compares leading theoretical approaches to address these challenges. We show that AI's development has brought escalating ethical challenges along with a maturation of frameworks proposed to address them. We map an arc from early monisms (e.g., deontology, consequentialism) to a variety of pluralist ethical frameworks (e.g., pluralistic deontology, augmented utilitarianism, moral foundation theory, and the agent-deed-consequence model) alongside pluralist governance regimes (e.g., principles from the Institute of Electrical and Electronics Engineers (IEEE), the United Nations Educational, Scientific and Cultural Organization (UNESCO), and the Asilomar AI principles). We find that pluralism is both normatively and operationally compelling: it mirrors the multidimensional problem space of AI ethics, guards against failures (e.g., reward hacking, emergency exceptions), supports legitimacy across diverse sociotechnical contexts, and coheres with extant principles of AI engineering and governance. Although pluralist models vary in structure and exhibit distinct limitations, when applied with due methodological care, each can furnish a valuable foundation for AI ethics.

artificial intelligence AI ethics value alignment black-box problem human-in-the-loop ethical pluralism policy guidelines

Genesis and History

On Artificial Intelligence

Artificial Intelligence (AI) refers to computational systems capable of behaviors that humans consider "intelligent," such as learning, reasoning, perception, and problem solving [1]. The field emerged in the mid-20th century, grounded in the aspiration to develop machines capable of emulating human cognitive functions. Early AI scholarship included Alan Turing's 'test' to distinguish between AI and human natural language responses, and a 1956 Dartmouth College workshop that coined the term "AI" [2][3]. John McCarthy, who led the Dartmouth workshop, proceeded "on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [4][5]. With

McCarthy's conjecture—now a goal—in mind, AI research grew markedly over the next seventy years, traversing multiple paradigms.

In the 1950s–1970s, researchers focused on symbolic AI using explicit, rule-based programs. This era emphasized knowledge representation and formal logic (e.g., production rules, theorem provers), such as "expert systems" for diagnostics and decision support [6]. The 1980s–2000s saw the rise of statistical learning methods, including AI utilizing decision trees, support vector machines, and the first multi-layer neural networks. Probabilistic models (e.g., Bayesian networks) and ensemble methods broadened the field, enabling learning from uncertainty and improving generalization [7].

From the 2010s onward, deep learning with neural networks enabled breakthroughs in AI perception and natural language processing [8][9]. These gains were driven by algorithmic advances (e.g., recurrent neural nets, parallel processing), new vast datasets, and accelerated computing capabilities. Most recently, increasingly sophisticated neural architectures (from reinforcement learning systems to large language models, generative AI, and agentic AI) have begun to rival human performance in healthcare, finance, law, and academia, albeit for narrowly defined tasks [10][11][12][13][14].

Terminological Housekeeping

AI systems have evolved substantially. The earliest symbolic AI encoded knowledge as explicit statements built from logic gates (e.g., "if-then", "and", "or"), and was engineered to apply logical inferences to derive reliable conclusions [9]. AI using decision trees modeled choices not as logic gates, but as hierarchical splits mapping features to predicted outcomes [8][9]. Related AI utilizing support vector machines identified boundaries to split data into groups, and classified data by margin width in a defined feature space [8][9].

Earlier AI systems employing Bayesian probabilistic models were able to represent uncertainty with explicit probabilities. These models updated credences from input data, thereby engaging in a rudimentary form of machine learning [8][9]. Machine learning (ML) refers to systems that are able to learn and fine-tune without being explicitly programmed to do so. ML has become increasingly sophisticated and widespread. AI ensemble systems group and combine ML AI systems that 'boost' or 'stack' each other to improve ML robustness, capacity, and accuracy [8][9].

AI systems utilizing neural networks are comprised of layered function approximators constructed by nodes in an interconnected network. At large scale with many net layers, large datasets, and parallel processing, this is referred to as deep learning [8][9][15]. AI systems using natural language processing (NLP) use algorithms to parse, interpret, and generate human language. Large language models (LLMs) are a class of NLP models that use deep neural networks to execute diverse language functions (such as ChatGPT, Claude, and Gemini).

Generative AI is a broad term for all AI models that synthesize, create, or generate new content such as text, images, audio, or code. Agentic AI denotes ecosystems of AI systems that collectively plan, utilize tools, execute

functions, and share memory, which are characteristically able to function without human prompting or oversight [16][17]. They are understood as autonomous or semi-autonomous AI ensembles [17].

AI capacity, architecture, and conceptualization have advanced across symbolic, statistical, and neural paradigms. While the term covers many systems and capabilities, in this entry, AI refers to a broad class of contemporary neural network-based systems, centered on LLMs, generative AI, and agentic AI. While different use cases carry different challenges, this entry consolidates the field's central issues and catalogs ethical frameworks marshaled to address them.

Structure of the Entry

With the terminological housekeeping now in hand, we start by articulating the rationale for AI ethics. Second, we trace the historical development of leading theories of AI ethics. Third, we compare these theories in the context of contemporary AI systems, highlighting the strengths of pluralist frameworks.

We begin by briefly sketching the novel challenges. Advances in AI have introduced new ethical challenges and magnified longstanding ones. These challenges, among the others outlined in Section 2, motivate the field. First, opacity: the rule-based symbolic systems of half a century ago exposed their reasoning, whereas contemporary generative and agentic models operate as black boxes, insofar as they rely on neural networks, complicating transparency and explainability [18]. Second, data provenance and bias: unlike earlier systems, modern AI models learn from vast, weakly governed datasets that are inscrutable and are prone to encoding biases, heightening auditability and fairness risks. By contrast, AI systems of the 1970s–1980s typically relied on highly curated knowledge bases [9]. Third, autonomy and control: agentic systems can pursue multi-step goals and can independently use external tools, yielding behaviors not explicitly programmed and increasing risks of unpredictability, misuse, and misalignment, concerns far less pronounced in early AI.

The risks are new, and the stakes are significant. AI systems are ubiquitous and are now embedded in many aspects of everyday life. AI systems sort and recommend content on social media, optimize logistics and supply chains, and power virtual assistants on our phones and computers [19]. These technologies influence outcomes of varying magnitude, such as who gets a loan or a job interview, the manner in which resources are allocated, and the way in which people access information [19].

These concerns are not speculative. AI systems have already been found to discriminate in hiring and lending [20], autonomous vehicles have been involved in fatal accidents [21], and chatbots trained on poor data and lacking ethical guardrails have produced misrepresentational outputs [22][23], and even contributed to suicides [24]. AI technologies are being increasingly employed in policing, medicine, law, academia, and warfare. Given AI's rapid development, its permeation across the human experience, and the potentially high stakes of failure, a robust framework for practicing AI ethics is needed.

To guide the responsible design, training, and deployment of AI systems, we need a full-bodied, operationalizable, and explanatorily powerful AI ethics framework. Amid rapid and disorienting change, we need guidance. The

remainder of this entry further motivates the rationale for AI ethics, traces the historical emergence of the field, explores major ethical frameworks, illustrates a turn from ethical monism to ethical pluralism, and endorses that progression.

References

1. Russell, S.; Norvig, P. Artificial Intelligence: A Modern Approach, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2010.
2. Turing, A.M. Computing machinery and intelligence. Mind 1950, 59, 433–460.
3. Legg, S.; Hutter, M. Universal intelligence: A definition of machine intelligence. Minds Mach. 2007, 17, 391–444.
4. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Mag. 2006, 27, 12–14.
5. Moor, J.H. The Dartmouth College Artificial Intelligence Conference: The next fifty years. AI Mag. 2006, 27, 87–91.
6. Anderson, J.; Rainie, L. The Future of Well-Being in a Tech-Saturated World; Pew Research Center: Washington, DC, USA, 2018.
7. Minsky, M.; Papert, S. Perceptrons; MIT Press: Cambridge, MA, USA, 1969; pp. 1–292.
8. Pearl, J. Probabilistic Reasoning in Intelligent Systems; Morgan Kaufmann: San Mateo, CA, USA, 1988; 552p.
9. Nilsson, N.J. The Quest for Artificial Intelligence: A History of Ideas and Achievements; Cambridge University Press: Cambridge, UK, 2009; pp. 1–558.
10. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021, 596, 583–589.
11. Li, Y.; Choi, D.; Chung, J.; Kushman, N.; Schrittwieser, J.; Leblond, R.; Eccles, T.; Keeling, J.; Gimeno, F.; Lago, A.D.; et al. Competition-level code generation with AlphaCode. Science 2022, 378, 1092–1097.
12. Trinh, T.H.; Wu, Y.; Le, Q.V.; He, H.; Luong, T. Solving olympiad geometry without human demonstrations. Nature 2024, 625, 476–482.
13. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. Nat. Med. 2025, 31, 943–950.

14. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A survey on large language model-based autonomous agents. Front. Comput. Sci. 2024, 18, 186345.
15. Cowan, J.D.; Sharp, D.H. Neural nets and artificial intelligence. Daedalus 1988, 117, 85–121.
16. Hughes, L.; Dwivedi, Y.K.; Malik, T.; Shawosh, M.; Albashrawi, M.A.; Jeon, I.; Dutot, V.; Appanderanda, M.; Crick, T.; De', R.; et al. AI agents and agentic systems: A multi-expert analysis. J. Comput. Inf. Syst. 2025, 65, 489–517.
17. Sapkota, R.; Roumeliotis, K.I.; Karkee, M. AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. Inf. Fusion 2025, 126, 103599.
18. Coeckelbergh, M. AI Ethics; MIT Press: Cambridge, MA, USA, 2020.
19. Beer, P.; Mulder, R.H. The Effects of Technological Developments on Work and Their Implications for Continuous Vocational Education and Training: A Systematic Review. Front. Psychol. 2020, 11, 918.
20. Lambrecht, A.; Tucker, C.E. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manag. Sci. 2019, 65, 2966–2981.
21. Macrae, C. Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk. Risk Anal. 2022, 42, 1999–2025.
22. Christian, B. The Alignment Problem: Machine Learning and Human Values; W.W. Norton: New York, NY, USA, 2020.
23. Dung, L. Current cases of AI misalignment and their implications for future risks. Synthese 2023, 202, 138.
24. Payne, K. An AI Chatbot Pushed a Teen to Kill Himself, a Lawsuit Against Its Creator Alleges. AP News. 25 October 2024. Available online: https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0 (accessed on 10 October 2025).