Aristotle's Ethical Deliberation and Virtue: Principles for Ethical AI

Alejandro Ricciardi

Colorado State University Global

HUM101: Critical Reasoning

Dr. Marc Meyer

August 4, 2024

Aristotle's Ethical Deliberation and Virtue: Principles for Ethical AI

The launch of the chatbot ChatGPT in November 2022 highlighted the potential of Artificial Intelligence (AI) to transform how people live and work. As AI systems become more advanced, their reasoning, decision-making, and generative capabilities raise profound ethical and moral questions. Incorporating Aristotle’s principles of ethical deliberation and virtue into AI systems' model training and development will ensure that these systems generate outputs and implementations aligned with human values, contributing positively to society's future.

Aristotle's Concept of Ethical Deliberation

Aristotle defines deliberation as a process of reaching a positive or good result by other means than just deduction. This process is defined as Aristotle's ethical deliberation and it is relevant to the debate about AI ethics, specifically regarding how AI systems can be trained to implement ethical reasoning when generating responses to prompts. Aristotle draws parallels between the process and geometrical analysis, where a mathematician works in reverse from the intended result. noting that "This 'seeking' aspect of deliberation is brought out in Aristotle's comparison of the deliberator to the geometer, who searches and analyzes by diagrams (EN 1112b20–24). Geometrical analysis is the method by which a mathematician works backwards from a desired result to find the elements that constitute that result" (Humphreys, n.d., 4.b).

Application to Modern AI Development

Aristotle's concept of ethical deliberation could be utilized to develop ethical AI systems by integrating these principles into AI model training. Generative AI, such as Large Language Models (LLMs), are autoregressive (AR) models (AWS, n.d.), meaning they generate outputs sequentially from an initial prompt. Therefore, Aristotle's deliberation concept could be used to incorporate ethical reasoning into AI system model training. This approach is particularly relevant given recent AI advancements, such as OpenAI's new AI model 'Strawberry'. According to Paul and Tang (2024) in their article "OpenAI Working on New Reasoning Technology Under Code Name 'Strawberry,'" this framework is capable of high-level reasoning (Tong & Paul, 2024).

Aristotle's Concept of Ethical Virtue

Another of Aristotle's concepts that could be implemented alongside and support the ethical deliberation process in AI development is his concept of ethical virtue or excellence. Aristotle defines it as "a deliberative disposition to take pleasure in certain activities, a mean between extreme states" (Humphreys, n.d., 4.a). In other words, Aristotle describes excellence as a (golden) mean between excessive and defective states of character. This concept of finding a 'golden mean' between extremes can be effectively implemented in AI systems, as the systems are inherently statistical models. This approach could help prevent scenarios like the one portrayed in the movie "I, Robot," where the AI system VIKI takes extreme measures, stating, "To protect humanity, some humans must be sacrificed. To ensure your future, some freedoms must be surrendered" (acorn8926, 2011, 0:00 – 0:55). Integrating Aristotle's concept of the golden mean with his concept of ethical deliberation into AI model training and implementation could help avoid generating unethical outputs by avoiding harmful extremes.

Challenges in Implementing Aristotle's Ethical Deliberation and Virtue in AI

In his writing about "Rhetoric" (Aristotle, n.d.), Aristotle notes: "Specific topics, on the other hand, are derived from propositions which are peculiar to each species or genus of things; there are, for example, propositions about Physics which can furnish neither enthymemes nor syllogisms about Ethics, and there are propositions concerned with Ethics which will be useless for furnishing conclusions about Physics; and the same holds good in all cases" (Aristotle, n.d., Chapter 2). This connects to the challenges of implementing Aristotle's ethical deliberation approach in AI systems, such as the difficulty of defining clear ethical concepts and the challenge of translating abstract ethical principles into forms that AI can process. Furthermore, the concept of endoxa (commonly held beliefs or reputable opinions) in dialectical reasoning translates, in AI system development, to the challenge of defining clear ethical goals and concepts that can be processed by AI systems. This highlights the complexity of implementing ethical reasoning in AI, where principles that apply in one domain may not be directly applicable or processable in another, much like Aristotle's observation about the specificity of propositions in different fields of study.

Conclusion

Aristotle's concepts of ethical deliberation and virtue offer a valuable approach to developing ethical AI systems. While implementation challenges exist, particularly in translating abstract principles into machine-readable formats, these ancient ideas remain remarkably relevant. As AI technologies advance, it is crucial to integrate these ethical concepts into the design of AI systems. This can help ensure that AI’s reasoning, decision-making, and generated outputs are always aligned with human values. Moving forward, AI developers need to embed these philosophical principles into the AI systems themselves to create technologies that are both powerful and ethical. By drawing on timeless philosophical wisdom and combining it with modern ethics, AI development can work towards a future where AI is safe to use and contributes positively to society.

References

Aristotle. (n.d.). Book I. *Rhetoric*. (W. Rhys Robert, Trans.). BOCC. acorn8926 (2011, May 10). *I, Robot-VICKI's monologue* [Video]. Youtube. https://www.youtube.com/watch?v=Np1A4AGpqSo&t=2s

AWS (n.d.). *What are Autoregressive Models?* Amazon. https://aws.amazon.com/what-is/autoregressive-models/

Humphreys, J. (n.d.). *Aristotle (384 B.C.E. - 322 B.C.E.).* Internet Encyclopedia of Philosophy. Retrieved Jun 23, 2024, from https://iep.utm.edu/aristotle/

Tong, A., Paul, K. (2024, July 15). *Exclusive: OpenAI working on new reasoning technology under code name 'Strawberry'*. Reuters. https://www.reuters.com/technology/artificial-intelligence/openai-working-new-reasoning-technology-under-code-name-strawberry-2024-07-12/