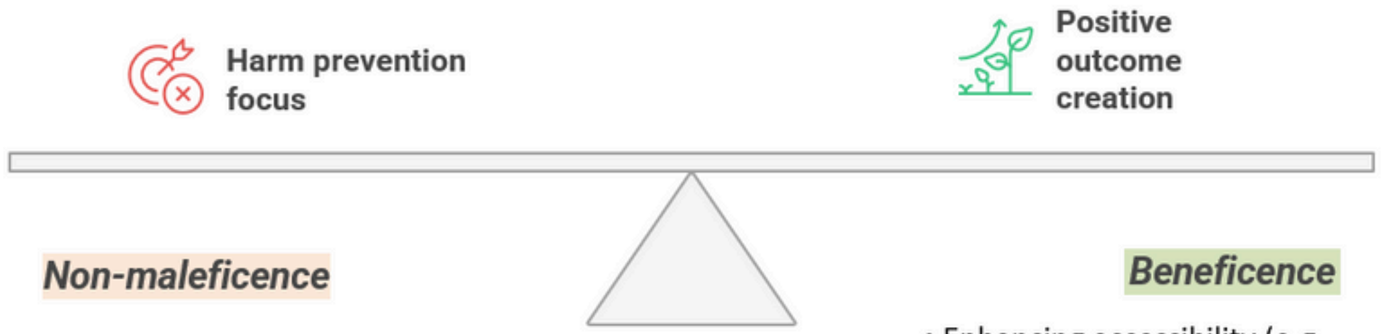
Responsible AI Starts with Ethics: A Step-by-Step Primer

Non-maleficence and Beneficence are core Ethical principles of AI

What is Non-maleficence and Beneficence in AI Ethics?

Non-maleficence ("do no harm") and beneficence ("do good") are complementary principles that must be balanced in the development and deployment of AI systems.



Non-maleficence

- Avoiding biased predictions that discriminate against groups
- Preventing safety failures in critical systems (e.g., healthcare, autonomous vehicles)
- Ensuring data privacy and avoiding misuse

Beneficence

- Enhancing accessibility (e.g., voice-to-text for the hearing impaired)
- Improving public services like education and healthcare
- Empowering marginalized communities through inclusive design

Non-maleficence & Beneficence Work in Harmony

Principle	*Focus*	*AI Example*
Non-maleficence	Avoiding harm	Ensuring an AI diagnostic tool does not misdiagnose or discriminate, protecting users from unintended negative consequences.
Beneficence	Promoting well-being	Designing AI to improve patient outcomes, such as early disease detection or personalized learning, actively enhancing lives and opportunities.

Key Points

- AI must be designed to maximize benefits while minimizing risks. For example, in healthcare, AI should support clinicians in making better decisions (beneficence) but also include safeguards to prevent errors or bias (non-maleficence).

- Regular assessment and stakeholder collaboration help ensure AI continues to do good and avoids harm as it evolves.

- These principles are not only about technical design but also about responsible governance, transparency, and accountability in AI's real-world use.

By integrating both principles, AI can be a force for positive change while protecting individuals and society from potential harms

AI System Lifecycle

Promote Social Good & Prevent ing Harm

Real World Examples

AI Context	*Applying Non-Maleficence*	*Applying Beneficence*	*Real-World Example*
Medical AI	Avoiding misdiagnosis in underrepresented groups	Improving diagnostic speed & accuracy	A major case involved an AI system making cheaper and less effective treatment recommendations for People of Color (POC) due to biased training data. The company identified and corrected the bias, retraining the model to ensure equitable care—thus preventing harm and improving outcomes.
Hiring AI	Preventing bias against women or minorities	Supporting fairer, skills-based hiring	AI hiring tools have been found to replicate gender and racial biases from historical data. Companies like Amazon discontinued biased recruitment AI after it was shown to disadvantage women, then shifted to more transparent, skills-based algorithms to promote fairer hiring practices.
Chatbots	Avoiding misinformation or harmful responses	Providing accessible mental health resources	A COPD management app's chatbot caused anxiety by suggesting possible lung cancer based on symptoms. After a complaint, the company revised the algorithm to avoid alarming messages, instead prompting users to consult a doctor—minimizing harm and improving supportive communication.

Continuation ...

• A pharmaceutical company's AI recommended less expensive treatments to POC patients due to biased data, potentially causing harm. After discovery, the company retrained the model for fairness and communicated transparently about the correction, directly addressing non-maleficence (avoiding harm) and beneficence (improving care for all).

• Real-world audits have shown AI hiring tools can perpetuate discrimination if trained on biased data. Amazon's discontinued recruitment tool is a notable example. The industry response has been to implement more transparent, skills-based AI systems to promote equitable hiring—actively supporting beneficence and non-maleficence.

• The BreathePro app for COPD patients caused unnecessary anxiety by suggesting a risk of lung cancer. After user feedback, the company revised the chatbot to provide more measured, supportive advice, demonstrating a correction to prevent harm and enhance user well-being.

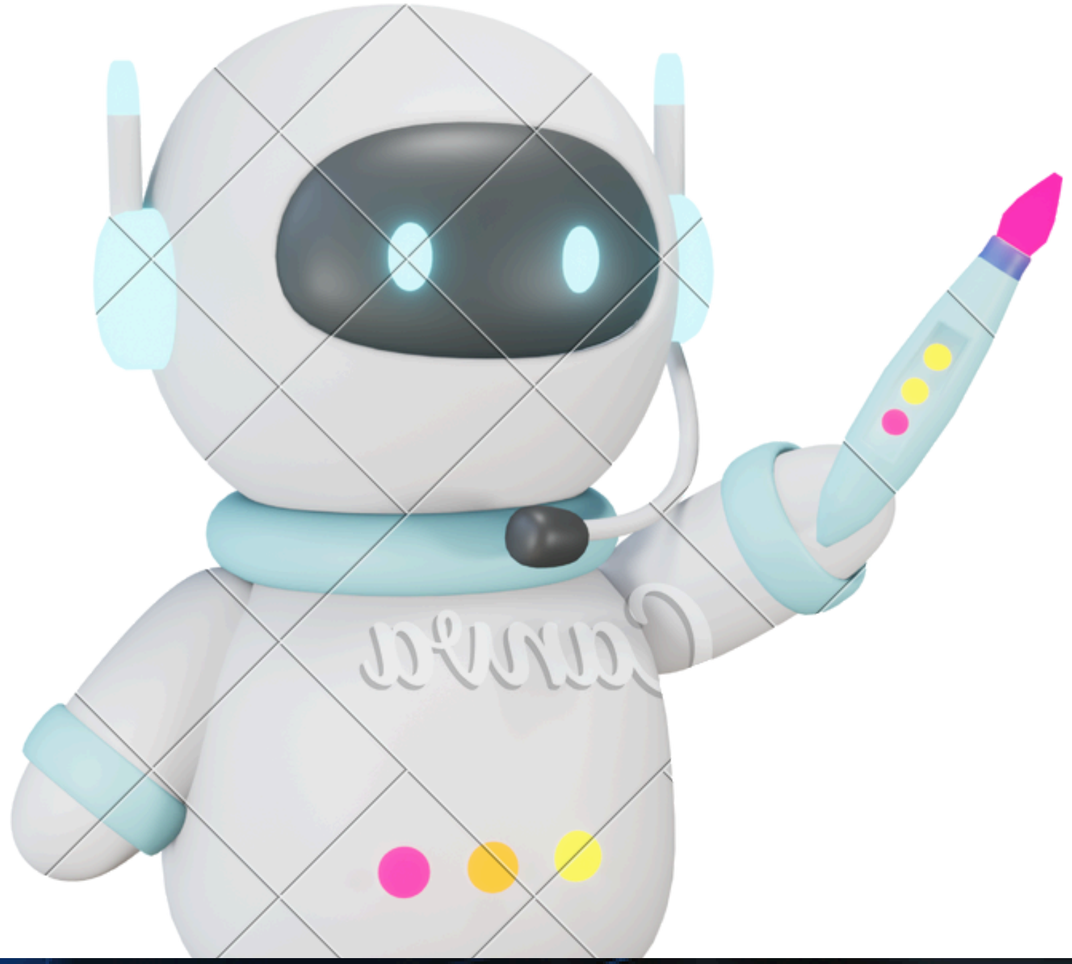
These cases illustrate how the principles of non-maleficence and beneficence are not only theoretical but are actively applied—and sometimes challenged—in real AI deployments, requiring ongoing vigilance, transparency, and corrective action to uphold ethical standards.

Deep Reading

- AI Educator Playbook – AI Principles (2025)
- An Introduction to the Ethics of AI in Education – SlideShare
- AI and Other Ethics – CAIML (PDF)
- AI Ethics Slides – WCET
- Components of an Ethical Framework for Artificial Intelligence in Education (2025, PDF)
- Artificial Intelligence in Education: Ethical Futures – SlideShare
- AI Ethics – Applying AI Ethics (AI for Good)

Reference Links

- AI Educator Playbook – AI Principles
- Components of an Ethical Framework for Artificial Intelligence in Education (PDF)
- AI and Other Ethics – CAIML (PDF)
- Artificial Intelligence in Education: Ethical Futures – SlideShare
- AI Ethics – Applying AI Ethics (AI for Good)
- An Introduction to the Ethics of AI in Education – SlideShare
- AI Ethics Slides – WCET



Thank You

Build powerful AI.
Build with ethics.
Build together.