CEI Center for Ethical Innovation

Design-for-Ethics in Artificial Intelligence

July 2024

Rowland Chen, S.M. M.I.T. Sloan School of Management
C.E.O., The Silicon Valley Laboratory Inc.
Executive Director, Center for Ethical Innovation
Adjunct Faculty, De Anza College, Cupertino, California
Former Visting Scientist, Carnegie Mellon University School of Computer Science

Image credit: 123rf.com

Red teams, today, are groups of artificial intelligence (AI) professionals who inspect and test the output of AI systems with an eye towards catching safety and ethics issues. The objective is to prevent AI that is harmful to humans from being unleashed on the world.

However, red teams come too late in the AI development lifecycle whether they are purely human or humans aided by AI chatbots. Self-regulation – one AI policing another AI – is a precarious situation. And red teams enter the software development cycle after the fact as is the case with OpenAI's Sora text-to-video generator. The teams attempt to retrofit ethics and safety once an AI has been coded and trained. Safeguards need to be built into the original code not testing safety into the code after it has been written.

Recall the ancient way of addressing product quality in manufacturing – design and build a product first and then check it for defects. It took decades for people to realize that engineers can design products for quality upfront as with Juran's Quality by Design approach popularized last century. Mindsets need to shift to a similar approach for AI ethics and safety by embedding ethical design principles throughout AI development.

Ethics embedded within AI involves making the technologies themselves intrinsically ethical. Required are the design of AI systems capable of understanding and adhering to moral principles autonomously. To date, building human traits into AI has proven to be a major, and perhaps insurmountable, technical challenge. Machine learning is possible. Machine conscience is elusive. AI researchers, designers, developers, and product managers can follow these five ethical AI design principles to embed safety and ethics from the start of their efforts:

1. Unbiased training data and accessed information
2. Algorithmic fairness of software designers and developers
3. Value alignment with cultural norms
4. Ethical reasoning
5. Autonomy and consent of humans

1. Unbiased Training Data and Accessed Information

The importance of unbiased training data and accessed information in embedding ethical innovation within artificial intelligence is an essential requirement for the development of AI systems that are fair, equitable, and reflective of a diverse society. Unbiased data ensures that AI algorithms produce output based on a balanced representation of the real world, avoiding the perpetuation of historical discrimination that can arise from skewed datasets. As AI continues to influence every aspect of our lives, the commitment to preventing bias becomes not just a technical necessity but a moral imperative to ensure ethical innovation.

2. Algorithmic Fairness of Software Designers and Developers

Algorithmic fairness starts with people. The goal is to develop AI that not only performs its intended tasks efficiently but does so in a manner that is unbiased. Joy Buolamwini (S.M. Algorithmic Bias, 2017), founder of the Algorithmic Justice League, is on a mission "to … prevent AI harm" Awareness building, education, motivation, and implementation of algorithmic fairness are required among all groups that are involved with ideation, design, development, and deployment of fair AI products and processes.

3. Value Alignment with Cultural Norms

Ensuring that AI objectives are in harmony with human values is an essential element for ethical integrity. Value alignment encompasses the establishment of goals that adhere to ethical standards while devising strategies to achieve these goals without causing unintended adverse effects. Value alignment is a significant challenge that requires ongoing dialogue among designers, technologists, ethicists, and the broader public. Alignment calls for a concerted effort to ensure AI software reflects human standards for acceptable decisions and behaviors.

4. Ethical Reasoning

Envisioning AI systems capable of ethical reasoning extends beyond programming decisions based on fixed ethical guidelines. The capability involves the development of AI that can assess various actions in new and complex situations to identify the most ethical path forward. These ambitious goals require AI to be endowed with a deep understanding of ethical principles and the ability to apply these principles across a spectrum of scenarios that are pre-trained and untrained. Crafting such systems demands a blend of technology, philosophy, and practical ethics, aiming to create AI that knows what is right and can discern with a degree of conscience the ethically best course of action in circumstances that are ambiguous, unprecedented, and untrained.

5. Autonomy and Consent of Humans

As AI systems evolve to operate with greater autonomy, it is necessary to ensure they respect *human* autonomy and the principle of consent. Designing AI that actively seeks and honors consent, particularly in applications where personal security and privacy are at stake, is critical. Artificial consent involves creating mechanisms within AI solutions that prevent manipulation, deceit, or coercion of users. Ensuring respect for human autonomy and consent requires a necessary design philosophy that safeguards human freedom in an increasingly interdependent world. And that includes dependence on AI.

Embedding ethics into AI is just one of several critical factors required for ethical innovation in AI. Others include the ethical use of AI and a computing machine's motivation for ethical behaviors. Granted, the design principles just described may pose insurmountable challenges today. A higher order of machine thinking is required, that could involve making breakthroughs in artificial creative intelligence and artificial consciousness.

Achieving this level of sophistication in AI requires a multidisciplinary approach, drawing from philosophy, psychology, cognitive science, cultural anthropology, and technology to design algorithms that can navigate ethical dilemmas using unbiased datasets. But that does not mean the red-team approach to AI safety is worthless. It is perhaps the best we can do for now. However, it is time to focus on building safety and ethics into AI from the get-go. The achievement of AI safety cannot go down the same path as manufacturing quality in the 20th century. AI red teams are too late.