AI User Code of Conduct

Daniel Hardman

2023-06-01

Abstract

Artificial intelligence is a tool, and like any tool it can be used well or badly. This document is a personal code of conduct — a public pledge governing the author's own use of AI, not a standard imposed on others. It sets out four commitments: to be truthful, and never use AI to deceive others about the humanness or identity of their counterparty; to respect social norms such as privacy, courtesy, and personal space; to accept responsibility for what an AI does under one's direction while still crediting its contribution; and to provide sensible safeguards and human backchannels against egregious automated mistakes. The aim is public accountability, and an invitation for others to write principles of their own.

I'm publishing this document to be publicly accountable and transparent about my intentions with respect to this important and ethically complex technology. The principles below seem important to me, but of course I encourage you to write your own document like this if you feel inspired. For ideas on code of conduct for AI builders*, I thought this open letter was insightful*.

Like all tools, AI confers power — for good or ill. To safeguard trust and human well being, I pledge to follow the guidelines below in my own use of AI.

1. Be truthful

I will not use AI to create content to deceive. I will not use an AI to interact with others in a way that violates reasonable expectations about the humanness, identity, or presence of their counterparty.

2. Respect boundaries

I will not use AI to violate social norms such as courtesy, privacy, personal space, and benefit of the doubt.

3. Accept responsibility but give credit

I accept responsibility for what an AI does under my direction. When I produce content with an AI's help, I will acknowledge the AI's contribution according to the reasonable expectations of my audience.

4. Provide safeguards and backchannels

I will take sensible measures to protect others from egregious mistakes made by AIs that I control or influence. I will allow people to escalate to me, short-circuiting layers of AI, when this is important.