AI Safety Fundamentals in Business, Medicine, and Social Media

Anonymous Author(s)
Affiliation
Address
email

Abstract

The responsible technology research community is dedicated to ensuring AI safety and developing solutions and governance mechanisms to mitigate risks for AI users. To achieve this, researchers are creating safety mechanisms such as benchmarks, red teaming, guardrails, policy interventions, and improving alignment to support better AI outcomes. Although these efforts aim to reduce exposure to unsafe AI outcomes, the challenge persists, particularly with the emergence of generative AI and agents. Many people have experienced unsafe AI outcomes, ranging from personal incidents to high-profile cases that receive news coverage. In response, a team of interdisciplinary researchers embarked on a project to investigate existing AI safety solutions and understand public concerns across various contexts, including medicine, social media, and business. After conducting a thorough analysis of the research landscape and reported unsafe AI experiences, the researchers developed an AI literacy campaign featuring a video and brochure. This campaign aims to educate the public about the use of AI in different domains, raise awareness about potential harms, and provide insights from experts. Additionally, it offers practical methods for promoting safer AI experiences, ultimately empowering users to navigate the benefits and risks of AI technology.

1 Education Submission Details

- **Target Audience:** General public, AI users
- **Expected Read/Watch Time:** 14 minute video; 5-minute brochure read
- **Material Description:** The AI Safety video begins with a short quiz of the viewer's recognition of real, and AI generated images, and the working definition of AI safety. The video contains brief interviews with individual responses to questions about AI use and concerns as it applies to business, medicine, and social media. For each AI application area, examples and case studies of AI risks are presented, along with definitions of the selected risk and solutions to address those risks. An accompanying brochure includes a timeline of AI accomplishments and questions for consideration for the use of AI in our selected contexts.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.