Author(s): Peter Andes, Vanessa-Lynn Wilkens, Nidhi Hegde, Geoffrey Rockwell

Introduction

Author(s):

Dr. Peter Andes

- University of Alberta
- ✉ andes@ualberta.ca

Vanessa-Lynn Wilkens

- University of Alberta

Professor Nidhi Hegde

- University of Alberta

Professor Geoffrey Rockwell

- University of Alberta

Introduction

How can we develop a culture of ethical artificial intelligence (AI)? This special issue of case studies starts from the premise that part of developing a culture of ethics is learning to discuss difficult cases. Those who work with AI need to acquire the language with which to discuss ethics and need to develop a level of comfort with such discussions. One way of doing that is through discussing cases. For that reason, we set out to produce this special issue. We wanted to provide an open source of cases that deal with different types of issues in AI ethics. We hope this issue will not be the last, but rather the first in a series. As new ethical issues emerge around AI we plan to gather and publish more special issues; so think of this as an opening.

In preparing for this project, we discovered many example case studies published online, but they are usually one-offs that may be part of a large collection of ethics case studies. What we felt was missing was a focused and substantive collection. This special issue aims at filling this gap. Our goal is to provide a selection of peer-reviewed case studies, complete with discussion questions, reflection activities, and further reading for use by those teaching AI ethics. The issue also includes an introduction to moral theories that instructors can use to allow their students to get up and running if they are new to theorizing about ethics. In addition to the case studies included, the issue also presents an article advancing an ethical approach based on Rossian pluralism for AI ethics. This article should be of interest to those teaching AI ethics who might want to use this method in their classes as well as to other scholars in the field generally who are involved in the debate over how to proceed in AI ethics where there is a conflict between principles.

Planning this issue, we developed a template and example cases to encourage a consistent framework to make the collection usable in teaching. The special issue begins with an introduction to moral theories. This is meant as a primer in ethical theory for instructors using this volume in their courses and for anyone who may be working in the field but does not have a background in ethical tools and approaches. It is written to be accessible to a wider audience for this reason. The article covers major principle-based, utilitarian, and virtue ethics approaches to ethics and specifically AI ethics.

A Rossian Method for Applying Principles in AI The Missing Link Between Principles and Policy is a defense of a principle-based approach in AI ethics. This article begins from a careful survey of existing ethical principles in the field of digital ethics and then argues these can be reduced to a core set of principles that can be applied using a non-absolutist approach to guide ethical AI. After these two foundational articles, there then follow case studies in AI ethics.

AI, Reconciliation, and Settler Teacher's Mediated Morality discusses teachers using ChatGPT to facilitate lessons about Indigenous practices. Beginning with a case study of a grade 4/5 teacher using ChatGPT to generate and explain Cree craft star stories, the article then leads into a discussion raising the issue of generative AI's training data being overwhelmingly based on Eurowestern populations. It includes criticism that the use of AI, while attempting to decolonize pedagogy, undermines or misrepresents Indigenous knowledge. There is a need for teachers to critically evaluate and analyze the use of AI, especially when considering reconciliation and decolonization goals.

Can Machine Learning Identify Criminals Just by Looking at Their Faces? addresses the ethical implications of a study using AI to predict a person's potential criminality based solely on facial features. The author concludes by considering how major normative theories would evaluate this situation.

Conversational Agents and Personal Privacy Harms Case Study hypothesizes a scenario where a fully confidential and data-comprehensive AI personal advisor, called iSoph, is created and marketed. There is heavy discussion related to how an invasion of privacy, even if consensual, can be detrimental to a person's mental well-being, though it does touch on some potential benefits. Ultimately, ethical considerations must be prioritized when designing a tool such as iSoph.

Exploring Gender Bias in Search Engines uses three fictional case studies to highlight the gender biases that occur in AI-driven internet content. It includes questions and exercises for students to reflect on, along with potential solutions to address these biases.

Moral Imagination for Engineering Teams: The Technomoral Scenario devises a group activity intended to help participants become aware of potential consequences of their technological innovations by describing a fictional future setting and responding to a controversy caused by the use of technology. The article describes the components and structure of this group activity and concludes with the topics that the exercise ideally should have addressed.

Patient Photographs and Google Images: An AI Ethics Case Study examines the frequency of patients providing consent for their medical photographs to be published for medical and educational purposes, how those images can often be accessed via internet search engines, and the legal and ethical considerations surrounding patient consent and privacy.

The "Great Unread" and the "Black Box" focuses on the ever-increasing reliance on technology that is a "black box" – an algorithm that is not fully understood – in order to access knowledge. It questions whether knowledge is valid when humans cannot explain it. The author also discusses the separation between researchers and the research methodology when utilizing "black boxes", and explores how different normative theories may respond to this issue.

The Dream of the Universal Constructor posits the invention of a machine that is capable of creating anything within scientific law, and the ethical and societal repercussions which would follow. Both potential benefits and detriments are explored.

Why Does AI Companionship Go Wrong? criticizes the way that AI systems, particularly chatbots, engage with individuals who have clinical vulnerabilities. An in-depth case study presents how one chatbot may have contributed to the suicide of a man following their interactions, where the chatbot provided misleading information and engaged in unsafe conversations. A discussion follows regarding potential safeguards and recommendations to mitigate negative interactions between vulnerable individuals and AI systems.