An AI Ethics Course Highlighting Explicit Ethical Agents

Dr. Nancy L. Green

THE UNIVERSITY of NORTH CAROLINA GREENSBORO

UNC GREENSBORO Department of Computer Science

Presented at 2021 Conference on AI, Ethics, and Society (AIES 2021)

Abstract

This is an experience report describing a pilot AI Ethics course for undergraduate computer science majors. In addition to teaching students about different ethical approaches and using them to analyze ethical issues, the course covered how ethics has been incorporated into the implementation of explicit ethical agents, and required students to implement an explicit ethical agent for a simple application. This report describes the course objectives and design, the topics covered, and a qualitative evaluation with suggestions for future offerings of the course.

Learning Objectives

I. Learn about some potential ethical impacts of AI systems,

II. Learn about some ethical theories, professional standards, and professional codes, and use them to analyze ethical issues in AI systems,

III. Learn about the difference between implicit and explicit ethical agents,

IV. Learn about some approaches to implementing explicit ethical agents, and

V. Implement an explicit ethical agent for a simple application and explain the ethical approach on which it is based.

Topics

- Introduction: Overview of issues in AI Ethics
- Ethical theories, codes, and standards
- Overview of design of ethical agents
- Bioethics of healthcare
- Implementation of explicit ethical agents for healthcare
- Ethics of warfare
- Implementation of explicit ethics agents for warfare
- Modeling emotion
- Formal logic based approaches
- Machine learning based system issues
- Social science approaches
- Cultural norms in ethics
- AI rebels (agents that can refuse to obey)

(See bibliography in full paper in AIES 2021 proceedings.)

Course Project

- Design and implement an explicit ethical agent
- Use logic programming language (Prolog) or IF-THEN rules so that ethical reasoning is transparent.
- Use any ethical approach covered in course. Students chose to use Ross' prima facie duties (fidelity, reparation, gratitude, justice, beneficence, nonmaleficence, and self-improvement), bioethics (beneficence, nonmaleficence, respect for autonomy), rule utilitarianism, etc.

Examples of agents:

- Daycare robot
- Robot elder companion
- Autonomous aerial delivery drone
- Robot bank guard
- Robot lifeguard
- Robot bartender

Ethical Agents

A central idea in the last three LOs is the distinction between implicit and explicit ethical agents (Moor 2006). An implicit ethical agent's actions are consistent with human ethical judgments but the agent has no explicit representation of ethical principles. However, an implicit ethical agent might encounter situations not anticipated by its designers. An **explicit ethical agent** could address such situations by reasoning about the ethical acceptability of its actions using an explicit representation of ethical principles (Scheutz 2017, Anderson & Anderson 2007). Thus, the fourth LO of the course is to study how explicit ethical agents have been implemented so far. The fifth LO is to apply that knowledge to the implementation of an explicit ethical agent.

Qualitative Evaluation and Suggestions for Future

- Topics: engagement in in-person class discussion good, positive comment in student evaluation ("interesting and thought-provoking")
- Format: change (due to COVID pandemic) to asynchronous class discussion less engaging than in-class discussion; in future remote learning, try synchronous video discussion
- Pandemic resulted in cancellation of planned outside speakers on data science and machine learning issues; more readings and discussion of implementation of ethical ML-based systems needed
- Project: Students without previous Prolog experience allowed to use simple IF-THEN rules for project implementation. In future need to provide more background in logic programming or more expressive IF-THEN rule language.
- Working assumption was that students enjoy exploring ethical approach via rapid prototyping. Next time increase value of hands-on by requiring each agent to be reimplemented using different ethical approaches and ask students to summarize the implications of each.