AI Ethics in Industry: A Research Framework

Authors

Ville Vakkuri, Kai-Kristian Kemell, Pekka Abrahamsson

Abstract

- AI systems are increasingly prevalent and their negative impacts on society are becoming apparent
- Early incidents have led to academic and public discussions on AI ethics
- Implementation of ethics in AI development is needed
- Little currently exists in the way of frameworks for practical implementation of AI ethics
- A research framework for implementing AI ethics in industrial settings is discussed in this paper
- The framework serves as a starting point for empirical studies into AI ethics
- The framework is still being developed further based on its practical utilization.

Introduction

Artificial Intelligence (AI) and Autonomous Systems (AS) have become increasingly prevalent in software development endeavors, changing the role of ethics in software development.

1. One key difference between conventional software systems and AI systems is that the idea of active users in the context of AI systems can be questioned. More often than not, individuals are simply objects for AI systems that they either perform actions upon or use for data collection purposes.
2. On the other hand, users of AI systems are usually organizations as opposed to individuals. This is problematic in terms of consent, not least because one may not even be aware of being used for data collection purposes by an AI.

Proposed Framework

1. Despite increased attention to AI ethics, a gap remains between research and practice
2. There are few empirical studies on the topic, and the state of practice is not well understood
3. A framework is needed to bridge the gap and provide a foundation for empirical research
4. The proposed framework is based on existing conceptual research in AI ethics
5. The framework has been practically applied to collect data
6. The paper evaluates the framework's performance based on this practical application

Overview of the Paper

1. Background: The Current State of AI Ethics
2. Research Model
3. Empirical Utilization of the Framework

Background: The Current State of AI Ethics

So far, the focus has been on four main principles for AI ethics:
1. Transparency,
2. Accountability,
3. Responsibility, and
3. Fairness

However, not all four of these values are universally agreed to form the core of AI ethics and effectiveness of using values or principles to approach AI Ethics has been criticized in and of itself.

Meeting the Growing Demands for Ethical AI

On an international level, the EU began to draft its own AI ethics guidelines which were presented in April 2019 (AI HLEG 2019). Moreover, the IEEE P7000™ Standards Working Groups ISO has founded its own standardization subcommittee (ISO/IEC JTC 1/SC 42 artificial intelligence.). Finally, larger practitioner organizations have also presented their own guidelines concerning ethics in AI (e.g.,

1. Google guidelines (Pichai 2018)),
2. Intel's recommendations for public policy principles on AI (Rao 2017),
3. Microsoft's guidelines for conversational bots (Microsoft, 2018)).

Research Model

To categorize the field of AI ethics three categories have been presented:

(1) Ethics by Design (integrating ethics into system behavior);
(2) Ethics in Design (software development methods etc. supporting implementation of ethics); and
(3) Ethics for Design (standards etc. that ensure the integrity of developers and users).

In this model, we focus on the latter two categories

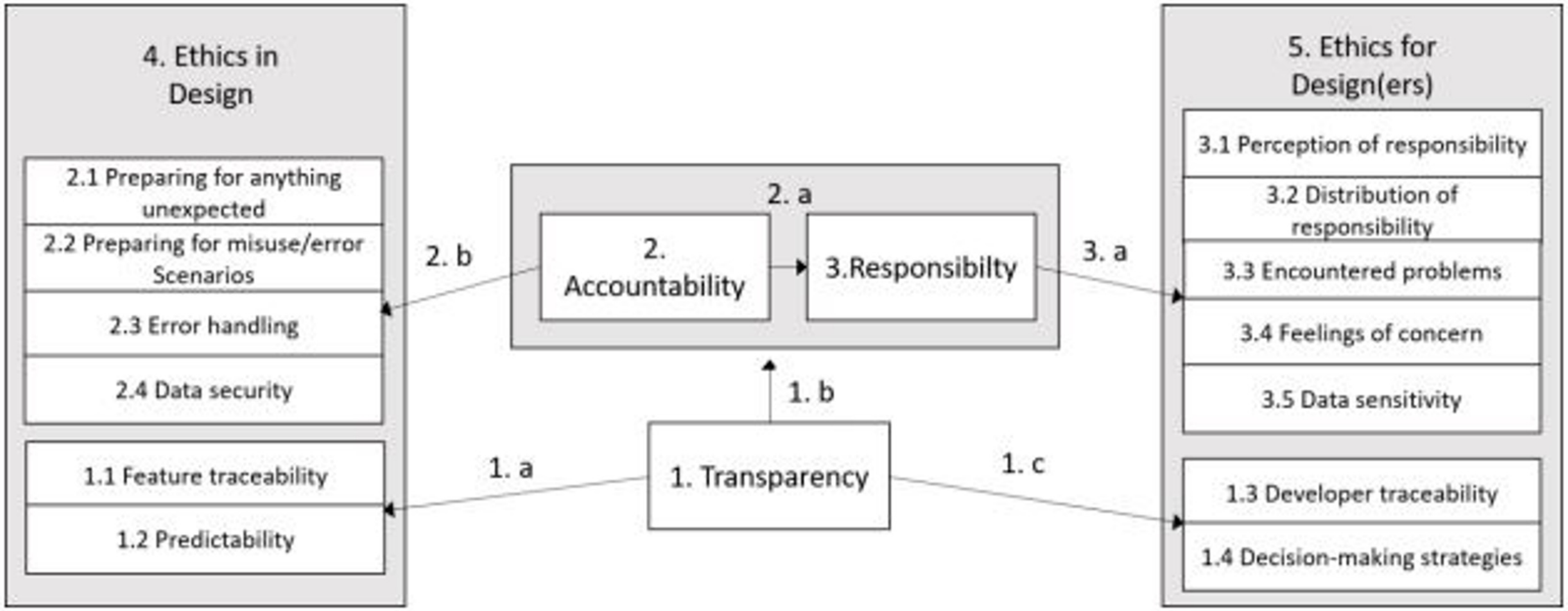

Figure 1 Research framework

The ART Model - Accountability

1. Accountability refers to determining who is accountable or liable for the decisions made by the AI.
2. Accountability is also defined to be the explanation and justification of one’s decisions and one’s actions to the relevant stakeholders.
3. In the context of this research framework, accountability is used not only in the context of systems, but also in a more general sense. Accountability is perceived through the concrete actions of the developers
4. Preparing for anything unexpected, Preparing for misuse/error scenarios, Error handling, Data security.

The ART Model - Responsibility

Responsibility is more focused on the internal processes of the developers not necessarily directly related to any one action. In order to act responsibly, one needs to understand the meanings of their actions.
In the research framework responsibility is perceived through the actions of the developers concerning :-

1. perception of responsibility
2. distribution of responsibility
3. encountered problems
4. feelings of concern
5. data sensitivity

The ART Model - Transparency

1. Transparency is a key ethical construct that is related to understanding AI systems.
2. In the research framework presented in this paper, we consider transparency not only in relation to AI systems but also in relation to AI systems development.
3. For the system to be considered transparent, feature traceability should be present, and the system should be predictable in its behavior.
4. For development to be considered transparent, the decision-making strategies of the endeavor should be clear, and decisions should be traceable back to individual developers.
5. Transparency also produces the possibility to assess accountability and responsibility in relation to both development and the system.

Empirical Utilization of the Framework

The research framework was utilized to carry out a multiple case study of three case companies. Each company was a software company developing AI solutions for the healthcare industry.

- The research framework, described in the preceding section,
- was utilized to construct the research instrument with which the data was collected.
- The questions prepared for the semi-structured interviews focused on the components of the framework.
- The interviews were recorded and the transcripts were analyzed for the empirical study
- Each transcript was first analyzed separately, after which the results of the analysis were compared across cases to find similarities

Empirical Results

The findings of the empirical study conducted using this framework were summarized into four Primary Empirical Conclusions (PECs).

● PEC1 Responsibility of developers and development is under-discussed
● PEC2 Developers recognize transparency as a goal, but it is not formally pursued
● PEC3 Developers feel accountable for error handling on programming level and have the means to deal with it
● PEC4 While the developers speculate potential socioethical impacts of the resulting system, they do not have means to address them.

Conclusions and Future work

1. Though the framework, as is, can be utilized for empirical studies, it should be complemented by the inclusion of some of the more recent AI ethics constructs such as fairness and trustworthiness to make it more current. Given that the framework was originally devised in late 2017, the discussion in the field of AI ethics has since then gone forward.
2. Aside from simply expanding the framework to include fairness and trustworthiness, we have plans to essentialize the framework by utilizing the Essence Theory of Software Engineering (Jacobson et al. 2012, Jacobson et al. 2017) in order to make it more relevant to practitioners.

Personal Views

AI ethics in industry is an important topic that requires attention and action from stakeholders. The development and implementation of ethical frameworks for AI can help ensure that AI systems are designed and used in ways that promote fairness, transparency, accountability, and privacy. Research in this area can help identify ethical challenges and provide guidance for the responsible development and deployment of AI technologies.

Thank You