Review of Ethically Aligned Design

by The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

Stephen Downes
National Research Council
February 27, 2018

What is this paper?

- Published by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- Purpose: advance a public discussion about how we can establish ethical and social implementations for intelligent and autonomous systems and technologies
  - aligning them to defined values and ethical principles
  - prioritize human well-being in a given cultural context
  - inspire the creation of Standards (IEEE P7000 series and beyond)
  - facilitate the emergence of national and global policies

See: http://www.downes.ca/post/67549

The Process

- The P7000 series was established to identify consensus on a broad range of ethical issues
- It's broken into a set of committees that each contribute to the report, eg:
  - Transparency of Autonomous Processes (7001)
  - Data Privacy Process (7002)
  - Standard on Child and Student Data Governance (7004)
  - Standard on Personal Data AI Agent (7006)
  - Etc.
- Comments for the current report (v23) are due March 18, 2018

See: https://ethicsinaction.ieee.org/

General Principles

- Embody the highest ideals of human beneficence as a superset of Human Rights.
- Prioritize benefits to humanity and the natural environment from the use of A/IS.
- Mitigate risks and negative impacts, including misuse.

Beneficence

From Wikipedia: “a concept in research ethics which states that researchers should have the welfare of the research participant as a goal of any clinical trial or other research study.” https://en.wikipedia.org/wiki/Beneficence_(ethics)

From SEP: “The term beneficence connotes acts of mercy, kindness, and charity. It is suggestive of altruism, love, humanity, and promoting the good of others.” https://plato.stanford.edu/entries/principle-beneficence/

General Principles

1. Human Rights – eg. Universal Declaration of Human Rights (1947)
   http://www.un.org/en/universal-declaration-human-rights/
2. Prioritizing Well-being – eg. OECD Guidelines of Measuring Well-being
   http://www.oecd.org/statistics/oecd-guidelines-on-measuring-subjective-well-being-9789264191655-en.htm
3. Accountability
4. Transparency
5. Awareness of A/IS Technology Misuse – eg. EU GDPR
   https://ec.europa.eu/info/law/law-topic/data-protection_en

- Prioritize benefits to humanity and the natural environment from the use of A/IS.
- Mitigate risks and negative impacts, including misuse.

Committees in EADv1 - Revised Content

- Embedding Values into Autonomous Systems
- Methodologies to Guide Ethical Research and Design
- Safety and Beneficence of Artificial General Intelligence
- Personal Data and Individual Access Control
- Reframing Autonomous Weapons Systems
- Economics/Humanitarian Ussies
- Law

New Committees in EADv2

- Affective Computing
- Policy
- Classical Ethics in A/IS
- Mixed Reality in ICT
- Well-Being

Classical Ethics in A/IS

- Draws from both philosophical and religious traditions, eg.
  - Secular traditions, such as utilitarian ism, virtue ethics, deontological ethics
  - Religious and culture-based such as Buddhism, Confucianism, Ubuntu, Shinto
- Critiques concepts such as good and evil, right and wrong, virtue and vice
- Review philosophical foundations that define autonomy and ontology
  - Can there be autonomous systems?
  - Can there be amoral systems?
- Address notions such as responsibility and accountability for decisions made by autonomous systems and other AI

(p. 193)

Interlude: Engineers Doing Ethics?

- Ethics is a specialized discipline requiring deep background
- As a human and social science, statistical methodology is important
  - Do ‘engineers’ constitute a significant representation of opinion?
- Ethics is a matter of value, not fact
  - There is no ‘fact of the matter’
  - There may be no consensus on ethical value
  - It may be that ethics itself is the wrong frame to govern A/IS
- The discussion of ethics is presented as a series of engineering ‘issues’
  - Is ethics reducible to a series of issues?
  - Can it be understood as a application of principles or rules?

Foundations

- Establishing foundations for morality, autonomy, intelligence
  - Three traditional economic divisions: individual, family, society
  - Questioning the disconnection of the autonomous individual from wider society
  - Awareness of the economic and political dimensions informing ethics
- Recommendations:
  - Keep in mind that machines do not comprehend the moral or legal rules they follow
  - Expand the definition of 'ethics' to include "the classical foundations of economy" (p.195)
- My comments
  - Both of these recommendations can be contested
  - Why economics and politics? Why not physics or biology? Or astrology?

Agents and Patients

- Ie., the distinction between moral agents and moral subjects
  - Distinction between “natural” self-organizing system
    - Especially with respect to genetic algorithms and evolutionary strategies
  - Vs artificial, non-self-organizing devices
- Autonomy in machines defines how they act and operate independently in certain contexts
  - “in some cases manifest(s) seemingly ethical and moral decisions, resulting for all intents and purposes in efficient and agreeable moral outcomes”
  - human traditions, on the other hand (manifests) as fundamentalism under the guise of morality (p. 196)
  - John Stuart Mill’s ethics “provides a detailed and informed foundation for defining autonomy”

Agents and Patients (2)

- Recommendations
  - Recommended that the discussion first consider free will, civil liberty, and society from a Millian perspective
- Comments
  - I am not convinced that the division between machine ‘agents’ and ‘subjects’ is as clearly defined as suggested here
    - In particular, even advanced algorithms are designed with objectives and background assumptions
  - I am concerned that the distinction cleaves the *responsibility* for an AI’s actions from the designer and places it on the machines
  - From an ethical perspective, therefore, so long as machines are *designed* and *owned*, they are not moral agents in and of themselves

Vocabulary

- Not everyone has the education and background to understand philosophical ethical terminology
- But philosophical theories are still useful
- Recommendation
  - Support groups raising awareness for social and ethics committees
  - Have philosophy scholars and ethicists present in no n-philosophy courses for A/IS technologists

Presenting Ethics to the Creators of A/IS

- Can classical ethics be used to produce meta-level orientations to data collection and data use in decision-making?
  - Key is to “embed ethics into engineering in a way that does not make ethics a servant, but instead a partner in the process” – a.k.a ‘ethics-in-practice’
  - Provide students with job aids & interatively increase complexity
  - Goal is to “provide students a means to use ethics in a manner analogous to how they are being taught to use engineering principles and tools.”
- Comment
  - Can ethics be represented as principles or tools?
  - Would engineers comprehend the moral or legal rules they follow?

Maintaining Human Autonomy

- Specifically: “the possibility of autonomous systems imitating, influencing, and then determining the norms of human autonomy”
  - Negation of independent human thinking
  - Machine’s definition of ‘who we are’ prevails
  - Example: Google’s autocomplete tool informs search patterns; search patterns used to define demographic categories
  - Example: use of person’s bio-information to support targeted advertising
- “Ultimately the behavior of algorithms rests solely in their design, and that design rests solely in the hands of those who designed them.” (p. 213)
  - Does the machine work *for* someone in particular? Or for particular groups?
- Recommendations
  - An ‘ethics by design’ methodology, and a ‘pointed and widely applied education curriculum’

Maintaining Human Autonomy

- Specifically: “the possibility of autonomous systems imitating, influencing, and then determining the norms of human autonomy”
  - Negation of independent human thinking
  - Machine’s definition of ‘who we are’ prevails
  - Examples: Google’s autocomplete informs search patterns, these then used to define demographic categories, or use of person’s bio-information to support targeted advertising
- “Ultimately the behavior of algorithms rests solely in their design, and that design rests solely in the hands of those who designed them.” (p. 213)
  - Does the machine work *for* someone in particular? Or for particular groups?
- Recommendations
  - An ‘ethics by design’ methodology, and a ‘pointed and widely applied education curriculum’

Rule-Based Ethics in Programming

- Deontological and teleological ethics that are best suited to simple moral machines
  - In deontology, *duty* is the point of departure
  - Duty can be translated into rules
    - For example, “Thou shalt not lie”
    - For example, Kant’s categorical imperative
  - Machines can follow simple rules
    - Inference engines are used to deduce new rules where necessary
    - These machines, though, work only in closed-world scenarios
  - In teleological models, the *consequences* of an action are assessed

Personal Data & Individual Access Control

- Automated and intelligent systems require data to support learning and decision-making, and this is often personal data (a.k.a. PII)
- Ethical considerations (p. 83):
  - What rights do people have to prevent information form being shared?
  - Individuals lack clarity about how to access, organize and share their own data
  - There's a need to "prioritize and include individuals" in the data processes
  - The right to have and provide informed consent
  - Essential need to "enable individuals to curate their identities and manage the ethical implications of their data"
  - Need to enable voluntary, rather than required, data collection from customers and citizens

Personal Data & Individual Access Control

- Automated and intelligent systems require data to support learning and decision-making, and this is often personal data (a.k.a. PII)
- Ethical considerations (p. 83):
  - What rights do people have to prevent information form being shared?
  - Individuals lack clarity about how to access, organize and share their own data
  - There's a need to "prioritize and include individuals" in the data processes
  - The right to have and provide informed consent
  - Essential need to "enable individuals to curate their identities and manage the ethical implications of their data"
  - Need to enable voluntary, rather than required, data collection from customers and citizens

Personal Data – Themes

- Digital Personas
- Regional Jurisdiction
- Agency and Control
- Transparency and Access
- Symmetry and Consent

These again are subdivided into issues

P7002, Data Privacy Process

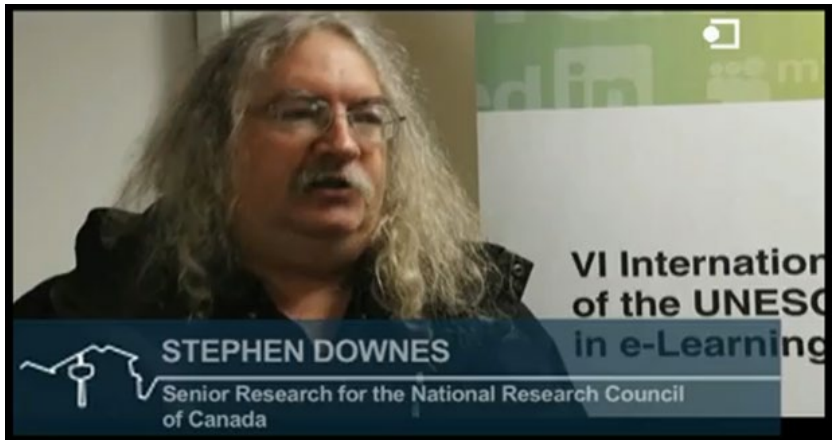
P7004, Standard on Child and Student Data Governance

P7005, Standard on Employer Data Governance

P7006, Standard for Personal Data Artificial Intelligence (AI) Agent

Digital Personas

- Digital identities and personals function differently in real and digital life
  - Behaviours considered a norm in real life and not so in digital life & vice versa
  - Eg. consider how we would use companion robots, autonomous vehicles
- Recommendations:
  - “policies, protections, and practices must provide all individuals the same agency and control over their digital personas and identity they exercise in their realworld iterations.”
  - “Individuals should have access to trusted identity verification services to validate, prove, and support the context-specific use of their identity.”
    - This includes data-verification services in sectors such as banking, government, telecommunications



Stephen Downes
National Research Council Canada
http://www.downes.ca https://www.nrc-cnrc.gc.ca/