Ethical Tensions

Roberto V. Zicari

Z-Inspection® Initiative

http://z-inspection.org

Graduate School of Data Science, SNU, Seoul

Structure of the Lesson

- *Ethical Tensions*
  - Definition of Ethical Tensions
  - Catalog of Ethical Tensions
  - Classification of Ethical Dilemmas
  - Trade Offs

Identify Tensions and Trade-offs

- **Tensions** may arise **between ethical principles, for which there is no fixed solution**.

- "In line with the EU fundamental commitment to democratic engagement, due process and open political participation, *methods of accountable deliberation to deal with such tensions should be established.* "

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

Reading

Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.*

Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), London. Nuffield Foundation.

Chapter 4.

*Exploring and addressing tensions. . . .*Page 19

Reading

- **How to Assess Trustworthy AI in Practice.** Roberto V. Zicari, Julia Amann, Frédérick Bruneault, Megan Coffee, Boris Düdder, Eleanore Hickman, Alessio Gallucci, Thomas Krendl Gilbert, Thilo Hagendorff, Irmhild van Halem, Elisabeth Hildt, Sune Holm, Georgios Kararigas, Pedro Kringen, Vince I. Madai, Emilie Wiinblad Mathez, Jesmin Jahan Tithi, Dennis Vetter, Magnus Westerlund, Renee Wurth

  arXiv:2206.09887 [cs.CY]. **[v2]** Tue, 28 Jun 2022

- **Chapter 2.5** Identify value conflicts and trade-offs

Ethical Tension: Definition

- “We use the umbrella term ‘tension’ to refer to different ways in which values can be in conflict, some more fundamentally than others.”

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* NuffieldFoundation.

Ethical Tensions

- “When we talk about tensions between values, we mean tensions between the pursuit **of different values in technological applications** rather than an abstract tension between the values themselves.“

- Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* NuffieldFoundation

Ethical Tensions

- This is related to our methodological choice to assess technological systems using socio-technical scenarios.
- We are not looking for abstract tensions between general principles, **but situations where highly valued principles are in conflict in a specific technological setting.**
- Of particular relevance are conflicts where norms, values or principles are mutually exclusive and thus cannot all be materialized or not all be materialized at the same level or with the same priority.

Ethical Tensions

- The ethical tension is most emphatically embedded in a technological device when highly praised norms or values are in conflict in the device itself or in its social use, and when a choice (at the design stage, or the deployment stage, or the use assessment stage) must be made.
- The whole methodology of the Z-inspection® leans towards such an ethical assessment of AI systems.

Specifying the most relevant ethical tensions

- After having reflected on the various norms, values and ethical principles that play a role in the technological application, the next step in the assessment process consists in specifying the most relevant ethical tensions. In general, at this step, central tensions between two or more relevant aspects were identified, with the focus on tensions between two norms, values or principles.
- *The task is to describe these tensions in open language.*
- For example, in a medical use case, examples could be: a tension between quality of services and autonomy; or between upholding standards and prevention of harm; or between efficiency and autonomy.

We use a Catalog of predefined Ethical Tensions

- To help the process, especially as a help to experts who might have not sufficient knowledge in ethics, we used a sample of catalog of predefined ethical tensions.

- We have chosen the catalog defined by the Nuffield Foundations (Whittlestone et al., 2019)

- Source: *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Catalogue of Examples of Tensions

From [1]:

- ***Accuracy vs. Fairness***
- ***Accuracy vs. Explainability***
- ***Privacy vs. Transparency***
- ***Quality of services vs. Privacy***
- ***Personalisation vs. Solidarity***
- ***Convenience vs. Dignity***
- ***Efficiency vs. Safety and Sustainability***
- ***Satisfaction of Preferences vs. Equality***

[1] Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Benefits

- This is another positive aspect of using socio-technical scenarios, since participants do not need to be fully trained in theoretical ethics to be involved in the debate.
- Ethical experts in the group can provide some expertise regarding the theoretical aspects of the discussion, while participating in the assessments of the socio-technical scenarios.

Accuracy *versus* fairness

- **Accuracy *versus* fairness**:

an algorithm which is most *accurate* on average may systematically *discriminate* against a specific *minority*.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Accuracy *versus* fairness: Hypothetical illustration

- To assist in decisions about whether to release defendants on bail or to grant parole, a jurisdiction adopts an algorithm that estimates the 'recidivism risk' of criminal defendants, i.e. their likelihood of re-offending.

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Hypothetical illustration (cont.)

- Although it is *highly accurate on average*, it systematically *discriminates against black defendants*, because the ‘false positives’ – the rate of individuals classed as high risk who did not go on to reoffend – is almost twice as high for black as for white defendants.

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Hypothetical illustration (cont.)

- Since the inner workings of the algorithm is a trade secret of the company that produced it (and in any case is too complex for any individual to understand), the defendants have little to no recourse to challenging the verdict that have huge consequences on their lives.

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Accuracy *versus* explainability

- **Accuracy *versus* explainability**:

the most accurate algorithms may be based on complex methods (such as deep learning), the internal logic of which its developers or users do not fully understand.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Quality of services *versus* privacy

- **Quality of services *versus* privacy**:

using personal data may improve public services by tailoring them based on personal characteristics or demographics, but compromise personal privacy because of high data demands.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Quality of services versus privacy: Hypothetical illustration

- A cash-strapped public hospital gives a private company access to patient data (scans, behaviours, and medical history) in exchange for implementing a machine learning algorithm that vastly improves doctors’ ability to diagnose dangerous conditions quickly and safely. The algorithm will only be successful if the data is plentiful and transferable, which makes it hard to predict how the data will be used in advance, and hard to guarantee privacy and to ensure meaningful consent for patients.
- Source: Whittlestone, J et al (2019)

Privacy *versus* transparency

- **Privacy *versus* transparency**:

the need to respect privacy or intellectual property may make it difficult to provide fully satisfying information about an algorithm or the data on which it was trained.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Personalisation *versus* solidarity

- **Personalisation *versus* solidarity**:

increasing personalisation of services and information may bring economic and individual benefits, but risks creating or furthering divisions and undermining community solidarity.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Personalisation *versus* solidarity: Hypothetical illustration

- A company markets a new personalised insurance scheme, using an algorithm trained on rich datasets that can differentiate between people in ways that are so fine-grained as to forecast effectively their future medical, educational, and care needs.The company is thus able to offer fully individualised treatment, better suited to personal needs and preferences.
- The success of this scheme leads to the weakening of publicly funded services because the advantaged individuals no longer see reasons to support the ones with greater needs.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Convenience *versus* dignity

- **Convenience *versus* dignity**:

increasing automation and quantification could make lives more convenient, but risks undermining those unquantifiable values and skills that constitute human dignity and individuality.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Convenience *versus* dignity: Hypothetical illustration

- AI makes possible an all- purpose automated personal assistant that can translate between languages, find the answer to any scientific question in moments, and produce artwork or literature for the users' pleasure, among other things. Its users gain unprecedented access to the fruits of human civilization but they no longer need to acquire and refine these skills through regular practice and experimentation.
- These practices progressively become homogenised and ossified and their past diversity is now represented by a set menu of options ranked by convenience and popularity.

- Source: Whittlestone, J et al (2019)

Satisfaction of preferences *versus* equality

- **Satisfaction of preferences *versus* equality**:

automation and AI could invigorate industries and spearhead new technologies, but also exacerbate exclusion and poverty.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Identifying further tensions

„Thinking about tensions could also be enhanced by systematically considering different *ways* that *tensions are likely to arise.*“

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S.

(2019), *London.* Nuffield Foundation.

Identifying further tensions

Whittlestone et al. points to three axis for such considerations: **power, time and locus** as follows:

Identifying further tensions

- **Winners versus losers**. Tensions sometimes arise because the costs and benefits of ADA-based technologies are unequally distributed across different groups and communities.

- Source: Whittlestone, J et al (2019)

Identifying further tensions

- **Short term versus long term**. Tensions can arise because values or opportunities that can be enhanced by ADA-based technologies in the short term may compromise other values in the long term.
- Source: Whittlestone, J et al (2019)

Identifying further tensions

- **Local versus global**. Tensions may arise when applications that are defensible from a narrow or individualistic view produce negative externalities, exacerbating existing collective action problems or creating new ones.

- Source: Whittlestone, J et al (2019)

Power, Time, Locus or Scope

- This helps to flag the broader issues linked to *winners* versus *losers* i.e. **power,**
- *Short/Long term* i.e. **time**

and

- *Local versus Global* - i.e. the **locus or scope** - or the "good for whom?" question.

How to resolve ethical tensions

- Once the ethical tensions have been identified as part of the case study assessment, the next question is how – *if at all – these ethical tensions can be resolved*.
- Thus, the next step in the assessment process consists *in deciding which of the options available to choose*.
- For example, in the case of an identified ethical tension between efficiency and autonomy, whether to choose the option that respects autonomy or the option that safeguards efficiency.

Dilemmatic situations

- In general, dilemmatic situations are situations where a difficult choice between two options has to be made.
- There are different types of dilemmas, however.

Use a Classification of Ethical Tensions

Kinds of tensions , from: Whittlestone, J et al (2019)

- **True dilemma**, i.e. "a conflict between two or more duties, obligations, or values, both of which an agent would ordinarily have reason to pursue but cannot";
- **Dilemma in practice**, i.e. "the tension exists not inherently, but due to current technological capabilities and constraints, including the time and resources available for finding a solution";
- **False dilemmas**, i.e. "situations where there exists a third set of options beyond having to choose between two important values".

True dilemmas, dilemmas in practice, and false dilemmas

- In this step of the assessment, the distinction between true dilemmas, dilemmas in practice, and false dilemmas, as suggested by Whittlestone et al. proved to be very useful, as we have found in group sections it can be the area that is the least understood/known to those in domain area groups.

Genuine ethical dilemmas

- Genuine ethical dilemmas can be characterized as situations where a choice has to be made between two or more options, but, no matter how the decision is made, there will be negative moral consequences. Sinnott-Armstrong gives the following definition for a genuine moral dilemma:
- *"a situation where an agent has a* ***strong*** *moral obligation or requirement to adopt each of two alternatives, and neither is overridden, but the agent cannot adopt both alternatives."*
- Role in ensuring that societal stakeholders have a voice in determining which values to prioritize.

Genuine ethical dilemmas

- In genuine ethical dilemmas, there are strong reasons to do each of two things, but only one can be done. Dilemmatic situations involve a conflict of two norms. When a decision is made and one option chosen, no matter what the decision is, the option chosen will conflict with one of the norms, values or principles.
- Thus, with an ethical tension, for example, an ethical tension between efficiency and autonomy, a "true dilemma" implies that the group can either choose an option that is in conflict with autonomy, or an option that is in conflict with efficiency. There is no "good" option available that satisfies both norms, values or principles.

Genuine ethical dilemmas

- The interdisciplinary approach in the Z-inspection® can help surface such tensions from different perspectives. To anticipate and prioritize among principles and values of "true dilemmas" a similar interdisciplinary dialogue approach should be considered prior to determining whether and how to apply an AI use-case.

Dilemma in practice

- In contrast, a “dilemma in practice” is a dilemma where an ethical tension exists, but where this tension can be overcome in principle.
- The tension exists only for practical reasons and could be overcome, for example, if more money would be available or if a different technological approach would be chosen.

False dilemma

- A "false dilemma" is a situation where two options with conflicting norms, values, and principles exist, but where it is not the case that a forced choice between these two options has to be made.
- The reason is that another, third option exists that is less dilemmatic, i.e. that does not require a choice between two important norms, values and principles.
- In the example with an ethical tension between efficiency and autonomy, this could be an option that is efficient but does not risk autonomy because it makes an additional level of data protection available.

Describe an Ethical Tension with an open vocabulary

We present below an example of a *tension* identified when analyzing the emergency tool to detect cardiac arrest tool:

- **ID Ethical Tension (Open Vocabulary): ET4**
- Kind of tension: **True dilemma.**
- Trade-off: ***Fairness vs. Accuracy.***
  Description: The algorithm is accurate on average but may systematically discriminate against specific minorities of callers and/or dispatchers due to ethnic and gender bias in the training data.

Trade-offs and true dilemmas

- To the extent that we face a true dilemma between two values, any solution will require making trade-offs between those values: choosing to prioritise one value at the expense of another.

- Source: Whittlestone, J et al (2019)

Dilemmas in practice

- On the other hand, to the extent that we face a dilemma in practice, we lack the knowledge or tools to advance the conflicting values without sacrificing one or the other. In this case, trade-offs may or may not be inevitable, depending on how quickly and with what resources we need to implement a policy or a technology

Source: Whittlestone, J et al (2019)