Design of Ethics Tools for AI Developers

Carl Mörch
June 2020
Introduction

• I’m a postdoctoral researcher at the University of Montréal and Mila – Québec Artificial Intelligence Institute. I’m funded by the Observatory on the Societal Impacts of AI (OBVIA).

• I have worked on the design of an AI Ethics Tool: the Canada Protocol (under revision, AI in Medicine), accessible on ArXiv.
As you have seen throughout the past classes, AI Ethics is a very active and lively topic.

It’s a fairly recent field, that started burgeoning in the 60’s and has become more important over the past decade.

It is at the intersection of Technology and Data Ethics.

AI, as a field, raises different types of concerns and questions (Gibert, 2019)

- Risks related to a “stupid AI” (e.g. a bug, bad code)
- Risks related to an “evil” use of AI (e.g. mass manipulation or influence, cyberattacks)
- Risks related to a “superintelligent” AI (e.g. general AI)
- Risks related to a lack of Ethics (e.g. nothing was done or anticipated)
What is the goal of AI Ethics?

For Taddeo and Floridi, AI Ethics can be helpful to:
- Manage the risks related to the technology
- Responsibly manage the opportunities
- Explore future questions

In this perspective, AI Ethics tools can help:
- Mitigate risks by providing concrete solutions
- Create more sustainable AI systems (legally and ethically)
- Prevent adverse or unintended consequences

When could an AI Tool be useful?

1. IDEATION   
2. DESIGN    
3. PROTOTYPE  
4. TEST      
5. DEPLOYMENT

6. DISSEMINATION
What’s a tool?

• What’s the difference between:
  • A tool
  • A framework
  • An instrument
  • A method

• This could seem like a rhetorical question, but the term tool is quite vague and can encompass various strategies, shapes and forms

• Today we will see that AI Ethics might require all these mechanisms because fields of practices are very diverse.
Is there a need for AI Ethics Tools

Over 80 guidelines, ethical declarations were released in recent years (Jobin et al., 2019).

They tend to converge around five key principles: justice and fairness, non-maleficence, autonomy, responsibility and privacy.

Few of them provide tools to translate them into practice. This can be an issue as having access to converging principles “does not bring change in the design of algorithmic systems” (Floridi, 2019).

Another issue

• Principles are probably not enough.

• “Four characteristics of AI development suggest a principled approach may have limited impact on design and governance. Compared to medicine, AI development lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms.” (Mittelstadt, 2019)

• “These weaknesses in existing legal and professional accountability mechanisms for AI raises a difficult question: is it enough to define good intentions and hope for the best? Without complementary punitive mechanisms and governance bodies to step in when self-governance fails, a principled approach runs the risk of merely providing false assurances of ethical or trustworthy AI”

More vocabulary questions

First of all, what could be the goal of AI Ethics Tools? To encourage:

- Responsible AI?
- The design of ethical system?
- An accountable process?
- More liable projects?

As you saw in Westerlund’s previous class, the terminology greatly varies. If we don’t define a clear goal, there is a risk to aim too wide and to not be specific enough.

Tools to apply high level-principles

• It seems that “tools” could be helpful to apply high-level principles or serve as compliance strategies, such as the EU’s assessment list.

• This list was developed by the High-Level Expert Group, its aims are:
  • To help organizations develop “trustworthy” AI systems
  • To promote some best practices
  • To operationalize “the (EU’s) key requirements for ethical AI and offers guidance to implement them in practice”
An example: EU’s Assessment List

- Example of an item from Technical Robustness and Safety
  
  *Resilience to attack and security Q7*
  
- Did you assess potential forms of attacks to which the AI system could be vulnerable?
Another example

Intel AI Ethics Toolkit
Other initiatives

https://aiblindspot.media.mit.edu/
There is momentum

- Here’s an example of the recent Neural Information Processing Systems (NeurIPS) Conference’s Guidelines:

  - “11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work? Yes, no or only partially. In order to provide a balanced perspective, authors are required to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. **Authors should take care to discuss both positive and negative outcomes.** Indicate whether you believe the broader impact section was adequate.”

  - “12. Does the submission raise potential ethical concerns? This includes methods, applications, or data that create or reinforce unfair bias or that have a primary purpose of harm or injury. If so, please explain briefly. Yes or No. Explain if the submission might raise any potential ethical concern. Note that your rating should be independent of this. If the AC also shares this concern, dedicated reviewers with expertise at the intersection of ethics and ML will further review the submission. Your duty here is to flag only papers that might need this additional revision step.”
A typology

• Jessica Morley (Oxford, NHS) and Luciano Floridi (Oxford) conducted a systematic review of AI Ethics Tools (using ArXiv among others) for Machine Learning.

• They created a typology. Its goal is to “help practically-minded developers apply ethics at each stage of the Machine Learning development pipeline, and to signal to researchers where further work is needed.”

## A typology

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business and use-case</td>
<td>Problem/improvements are defined and use of AI is proposed</td>
</tr>
<tr>
<td>development</td>
<td></td>
</tr>
<tr>
<td>Design Phase</td>
<td>The business case is turned into design requirements for engineers</td>
</tr>
<tr>
<td>Training and test data</td>
<td>Initial data sets are obtained to train and test the model</td>
</tr>
<tr>
<td>procurement</td>
<td></td>
</tr>
<tr>
<td>Building</td>
<td>AI application is built</td>
</tr>
<tr>
<td>Testing</td>
<td>The system is tested</td>
</tr>
<tr>
<td>Deployment</td>
<td>When the AI system goes live</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Performance of the system is assessed</td>
</tr>
</tbody>
</table>

The Canada Protocol

- Here’s an example of what an AI Ethics tool can look like when applied to a specific field of practice.

- I come from the field of Psychology and I worked in a Suicide Prevention and Ethics lab (CRISE).

- AI and “Big Data” is increasingly used to
  - Monitor suicide risk (e.g. on social media)
  - Analyze large volumes of Electronic Health Records
Our goal

- We wanted to bridge a gap between tree distinct fields:
  - Mental Health
  - Ethics
  - Computer Sciences
What's in the Canada Protocol?

A checklist

Why? It is a very common tool in Healthcare, therefore designing a checklist could improve its usability in the targeted sector.

It is comprised of 38 items
Methodology

- We analyzed 14 documents to explore what could be the ethical challenges mentioned in the literature. We used:
  - Ethical Declarations (e.g. IEEE Intelligent Design)
  - Key documents on ethics & technologies in Healthcare

- We found 450 mentions of potential ethical issues and challenges

- After several stages of work with two researchers we ended up with a list of 45 items that was then submitted to a panel of 16 international experts (in AI, Ethics, Mental Health) for content validation (using the Delphi methodology).
<table>
<thead>
<tr>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objectives</strong></td>
</tr>
<tr>
<td><strong>Technology</strong></td>
</tr>
<tr>
<td><strong>Funding &amp; conflict of interest</strong></td>
</tr>
<tr>
<td><strong>Credentials</strong></td>
</tr>
<tr>
<td><strong>Target population</strong></td>
</tr>
<tr>
<td>**Evidence **</td>
</tr>
<tr>
<td><strong>Testing</strong></td>
</tr>
<tr>
<td><strong>Complaints</strong></td>
</tr>
<tr>
<td>PRIVACY &amp; TRANSPARENCY</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td><strong>Responsibility</strong></td>
</tr>
<tr>
<td><strong>Data collection</strong></td>
</tr>
<tr>
<td><strong>Accessibility</strong></td>
</tr>
<tr>
<td><strong>Informed consent</strong></td>
</tr>
<tr>
<td><strong>Consent withdrawal</strong></td>
</tr>
<tr>
<td><strong>Access to the data</strong></td>
</tr>
<tr>
<td><strong>Right to be forgotten</strong></td>
</tr>
<tr>
<td><strong>Minors</strong></td>
</tr>
<tr>
<td>SECURITY</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
<tr>
<td><strong>Embedded recording mechanism</strong></td>
</tr>
<tr>
<td><strong>Third-parties</strong></td>
</tr>
<tr>
<td><strong>Data protection</strong></td>
</tr>
<tr>
<td><strong>Audit trails</strong></td>
</tr>
<tr>
<td><strong>Autonomy</strong></td>
</tr>
<tr>
<td><strong>Moderation</strong></td>
</tr>
<tr>
<td>HEALTH-RELATED RISKS</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Type of care</strong> *</td>
</tr>
<tr>
<td><strong>Crisis &amp; contigency planning</strong> *</td>
</tr>
<tr>
<td><strong>Non-maleficience</strong> *</td>
</tr>
<tr>
<td><strong>Misuse</strong></td>
</tr>
<tr>
<td><strong>Emotions detection</strong> *</td>
</tr>
<tr>
<td><strong>Emotions control</strong> *</td>
</tr>
<tr>
<td><strong>Relationship</strong> *</td>
</tr>
<tr>
<td><strong>Public awareness</strong></td>
</tr>
<tr>
<td>Category</td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
<tr>
<td>Ethics</td>
</tr>
<tr>
<td>Exclusion &amp; discrimination</td>
</tr>
<tr>
<td>Stigmatization *</td>
</tr>
<tr>
<td>Detection</td>
</tr>
<tr>
<td>Data handling</td>
</tr>
<tr>
<td>Data selection</td>
</tr>
<tr>
<td>Data transformation</td>
</tr>
<tr>
<td>Other</td>
</tr>
</tbody>
</table>
Challenges

• There are some challenges when developing AI Ethics tools
• Professionals from different field do not always use the same terminology.

• When they do, there might be some misunderstanding (e.g. detection).

• Not everybody received a training in Ethics.

• We all have preconceived notions about others.
Ethical baseline: does it exist?

All the ethical guidelines and declarations give an indication of what could be an ethical “baseline”. But what could it be?

A minimal level of ethical requirements
A universal set of moral rules everyone should follow

But when this concept is applied to computer sciences, what form could it take? A list of technical requirements inspired by ethical standards? (e.g. Dean et al., 2016)

It seems difficult to establish a moral and universal baseline.

A few challenges

As promising as ethics tools can be, there are few misuses and challenges ahead of us.

For example, these tools and declarations are not exempt of criticism, including a potential contribution to ethical white washing.

“According to Philosopher Thomas Metzinger at Johannes Gutenberg University Mainz in Germany, big tech indulges in what he says is “ethics washing” and wields too much influence on proposed industry regulation around AI. According to Metzinger, big tech companies’ ethical debates on AI are red herrings to delay policy formation and regulation worldwide.”

EU guidelines

Ethics washing made in Europe

On Tuesday, the EU has published ethics guidelines for artificial intelligence. A member of the expert group that drew up the paper says: This is a case of ethical white-washing. VON THOMAS METZINGER

https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html
https://indvstrvs.com/artificial-intelligence-governance-needs-more-than-consensus/
Window dressing

• “Things that are of no real importance and are said or done in order to make an attractive effect. Example: How many of the candidate's policies are real intentions, and how many are just window dressing?” (Cambridge Dictionary)
Inclusion and Participation

How can we make sure that the tools are adapted to
The targeted field?
The people impacted by AI?

One strategy is to go above inclusion and promote active participation

Deliberations are a methodology that could serve this purpose
Example 1: We use it to improve UNESCO’s upcoming list of AI Ethics standards. We organize international deliberations.
Example 2: We will use it with Ethics Boards members across Québec
Conclusion

• AI Ethics declarations and tools could contribute to a more responsible development of AI.

• However, without strict regulation and enforcing legal rules, they could remain a form of “wishful thinking”.

• These tools also need to show more use-cases to demonstrate their validity, usefulness.
Thanks!

carl.morch@umontreal.ca