Using AI for decision support: some ethical issues

Ibo van de Poel

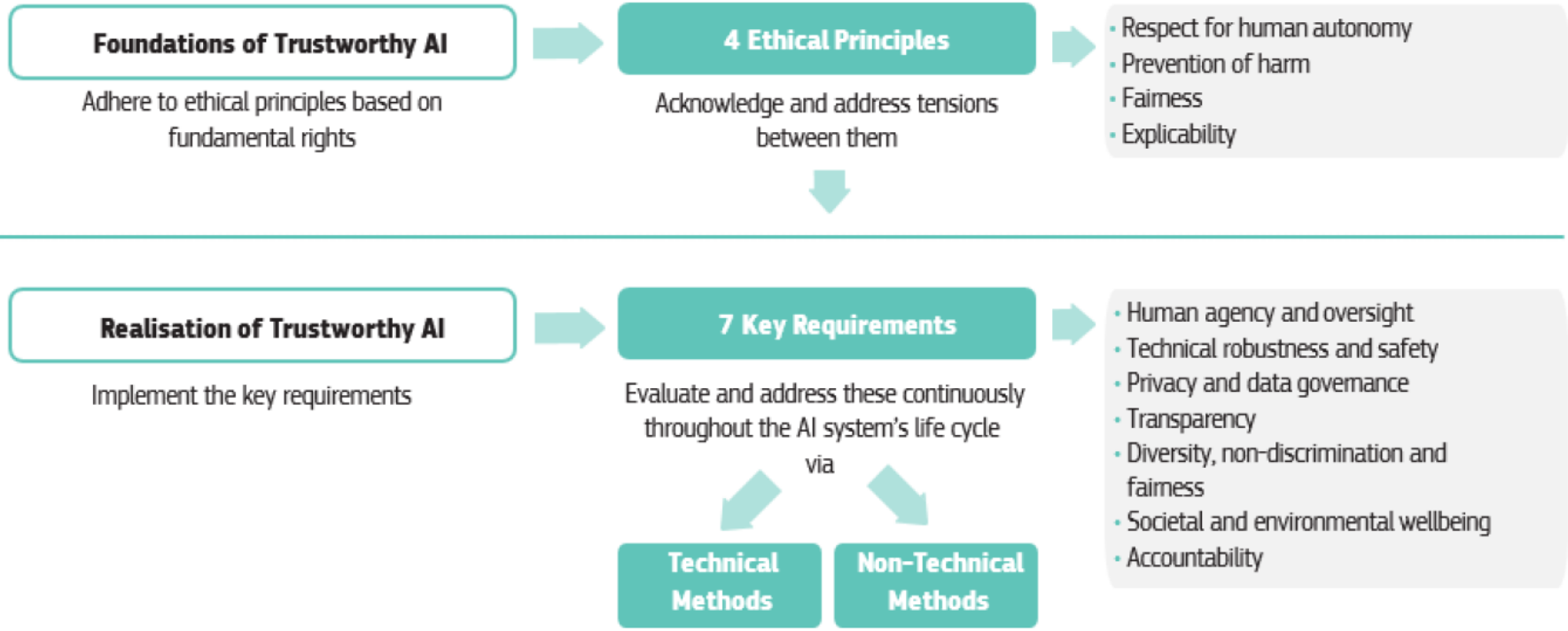
TU Delft

AI in publishing

- I will discuss some general ethical issues raised by AI but with an emphasis on a scientific publishing context

- I will focus on use of AI for *editorial decisions*

EU High-level Expert group on AI



Focus on three ethical concerns or values

- Bias and fairness
- Accountability
- Explainability
- (and a further ethical issue)

TU Delft

Bias

We want editorial decisions to be unbiased.

Two possible definitions:

1. Bias = certain elements in the data set are over-represented or get a higher weight
2. Bias = systematic and unfair discrimination against certain individuals or groups of individuals in favor of others

Possible causes of bias

- Design choices
- Training set
- Emergent
  - Learned by algorithm
  - New use
- Human decision maker

Existing bias in publishing?

If we use existing editorial decisions, what unintended biases might we get?

- Native versus non-native speakers?
- Fitting the existing paradigm versus non fitting the existing paradigm?
- Bias based on country?

Avoiding bias is not only important for doing justice to the individual author but also for scientific disciplines and for society, that relies on scientific results

TU Delft

Avoiding bias

Variables

Avoid sensitive variables

- But: proxy variables

Training set

Look for unbiased training set

- But how do we determine whether a training set is unbiased?

Fairness metrics

Use fairness metrics

- E.g., for dealing with emergent bias
- But there are various metrics which cannot all be optimized at the same time

Some questions that need to be asked

- What biases are clearly unacceptable?
- What biases are undesirable but perhaps not always unacceptable?
- What are current biases that we would like to avoid?
- What new biases might arise?
- What fairness metrics are most relevant for the context of publishing?

TU Delft

Focus on three ethical concerns or values

- Bias and fairness
- **Accountability**
- Explainability
- (and a further ethical issue)

Accountability

- Accountability = ability and willingness to account to others for one's decisions and actions
- Publishing company/editor has an accountability to authors but also more broadly to scientific community and to public
- Computers/AI cannot be accountable, because they lack (moral) agency
  - But they can be so designed as to help (or hinder) human accountability

Human in the loop

May be desirable, but is not always the solution:

- Too little: if humans do not have time, information, capabilities etc. to make decisions (epistemic enslavement)
- Too much: might be more important to have a clear owner of accountability than to have many humans in the loop (problem of many hands)

Meaningful human control

- Core decisions need to be made by humans in a meaningful way
  - i.e., enough time information etc.
  - Does not imply human in the loop, can also be human on the loop (operator level), or design decision made by humans

Conditions for meaningful human control

Tracing condition: Crucial decisions can be traced back to at least one human

Tracking condition The process should track the right kinds of reasons

Reason-responsiveness

- It is not inconceivable that a neural network is reason-responsive
- But if it cannot explicate these reasons to users, it might be of little use
  - Explainability

Questions to be asked

- Which humans should be accountable for what?
- How can we improve the overall accountability of the system? (rather than its parts)?
- How can AI help to improve accountability rather than erode it?

Focus on three ethical concerns or values

- Bias and fairness
- Accountability
- **Explainability**
- (and a further ethical issue)

Explainability

Results of AI may be unexplainable

Explainability can mean (many) different things; and explainable to whom?

Some explainability is needed to know whether we can trust/rely on the outcomes

Explainability and publishing AI

- For algorithm developers
  - To improve algorithm
  - E.g., to avoid spurious correlations and biases and to ensure fairness
- For editor (decision-maker)
  - To understand limitations and potential biases
  - To understand reasons and to communicate to author
- To author
  - Has right to know reasons for rejection (and should be good reasons)

TU Delft

Explainability and machine learning (ML)

- Machine learning techniques like reinforcement learning are prone to opaqueness
- There are various methods and techniques (being) developed to improve explainability
- However, often the focus is on causal explanations
- For editorial decisions, we need more than causal explainability; we need justification based on reasons

TU Delft

Questions to be asked

- What are the explanatory needs of the various users/stakeholders?
- What is required to serve these needs?
- How can we move beyond causal explanations towards justification based on reasons?

A further ethical issue

- Gaming the system
  - As soon as people know/understand how editorial AI system work they may try to game the system
  - Attempts to avoid this may be at tension with accountability and explainability!