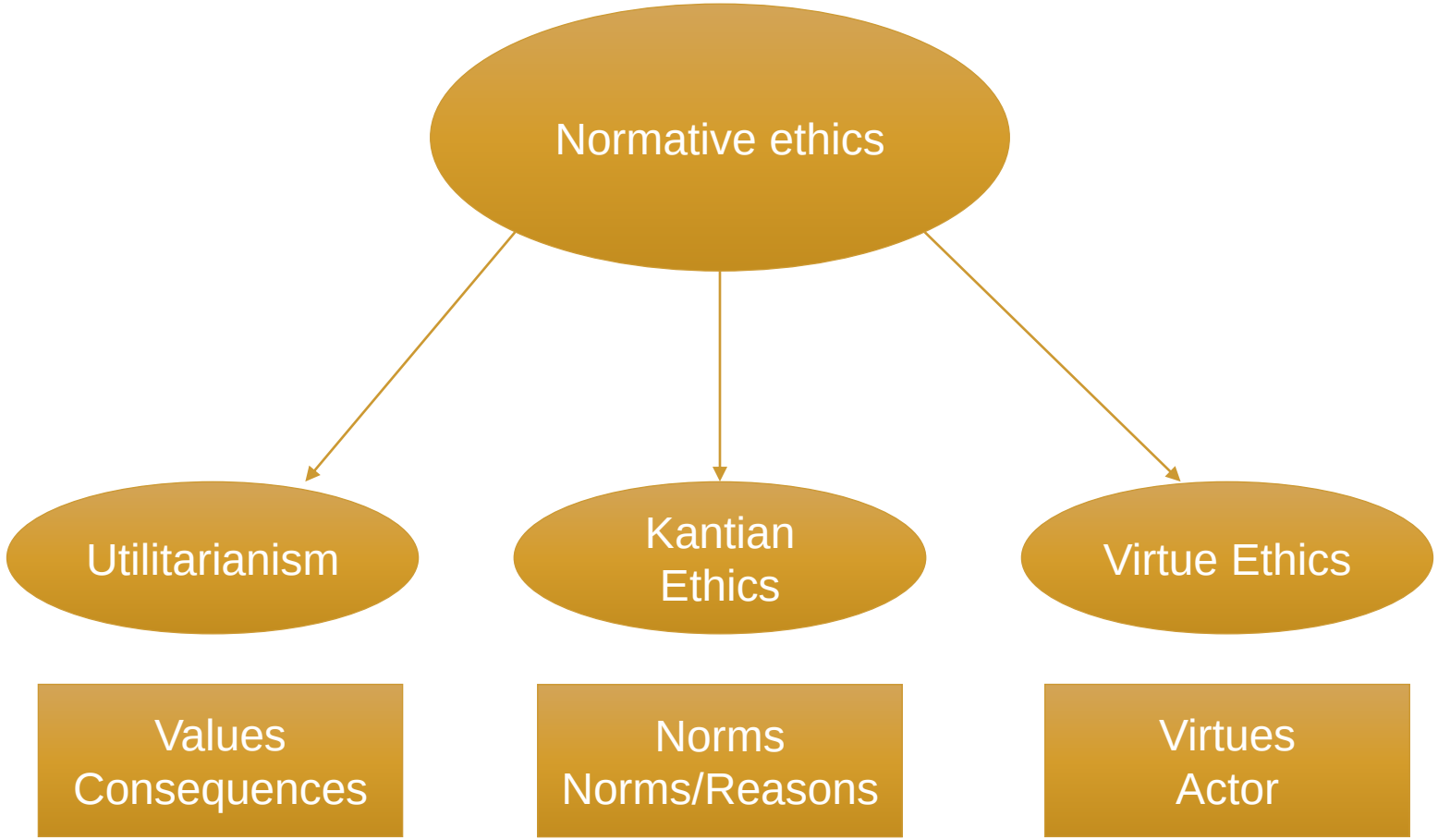
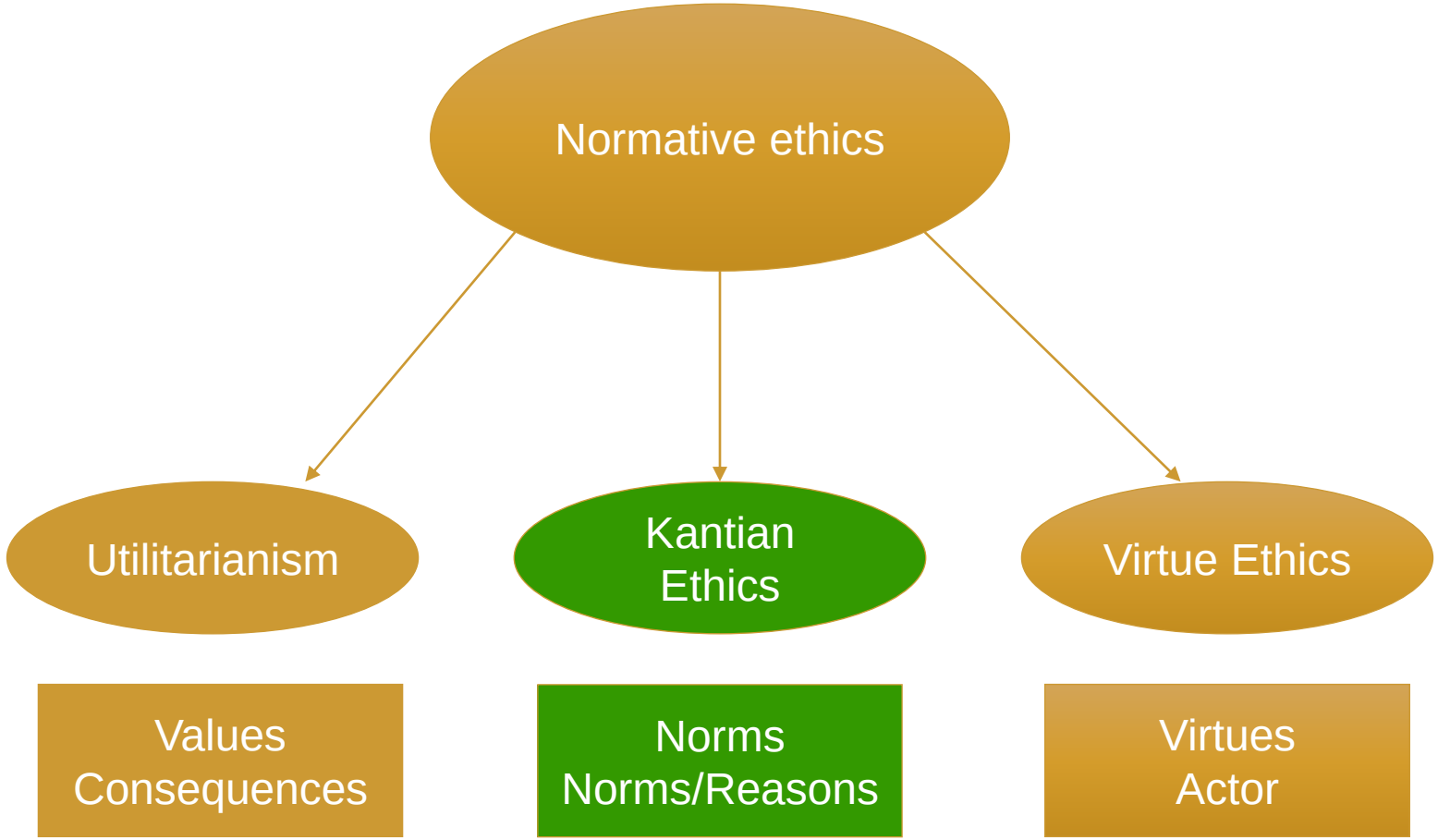
Module 3: Kantian Ethics (deontology) and explanation

Dr. Gert Meyers (TILT, Tilburg University)

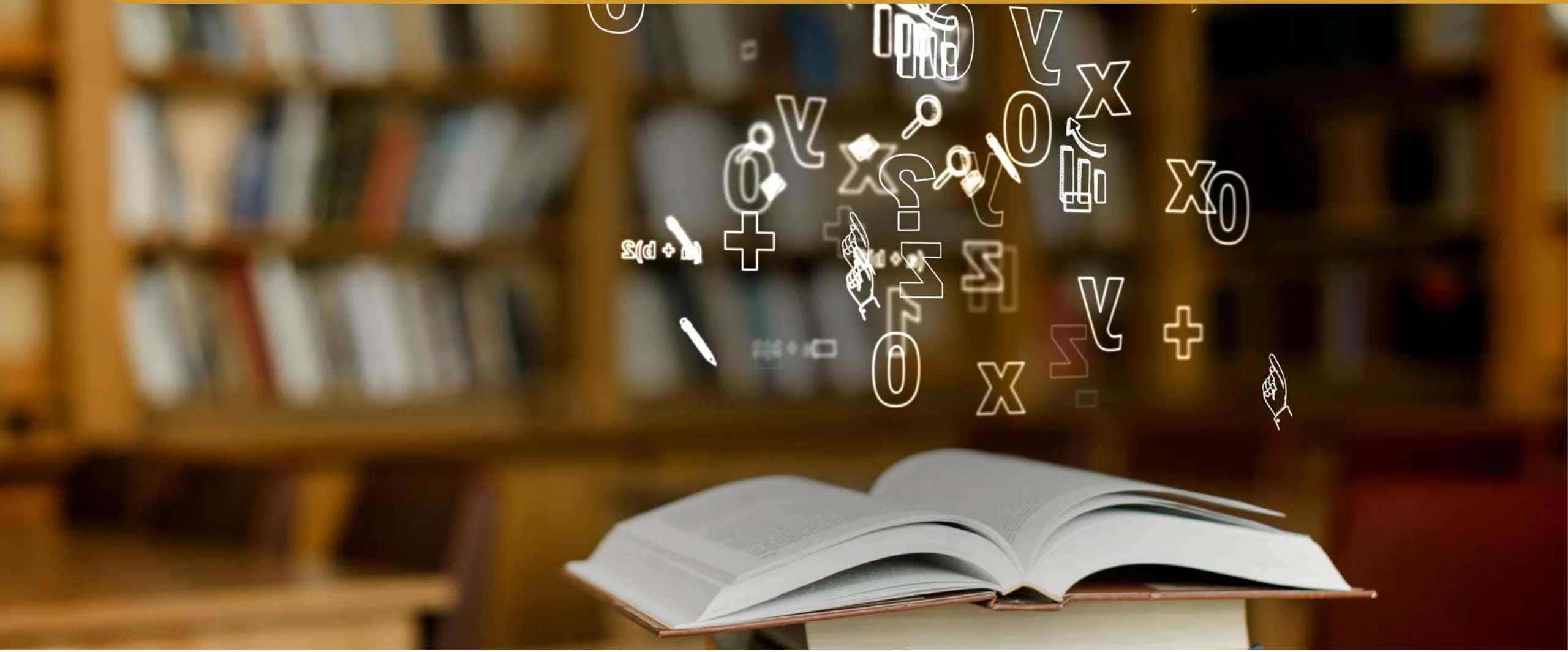
Ethical theories



Ethical theories



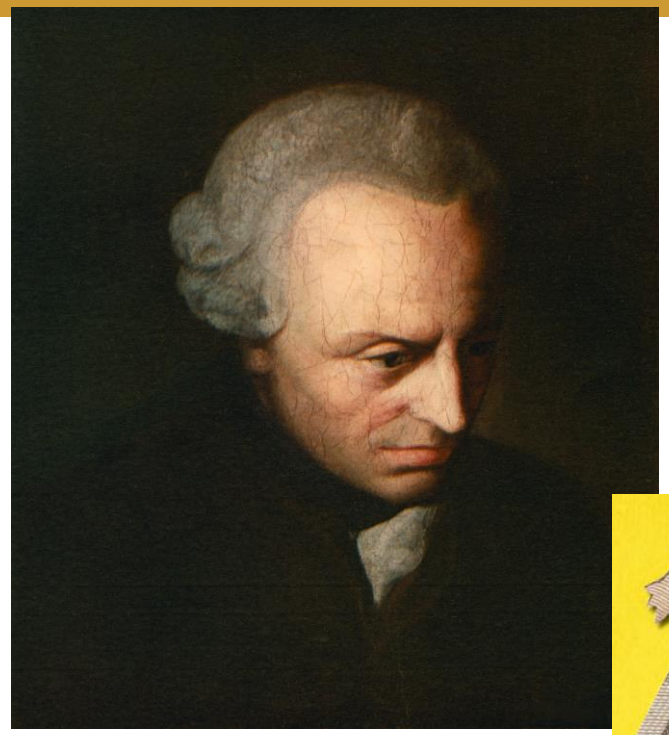
Questions?



Tilburg University

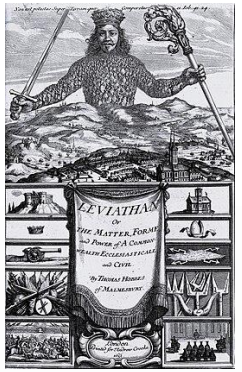
Immanuel Kant

- 1724-1804
- Renowned philosopher
- Major work: Critique of pure reason (1781)
- Foundations of the metaphyiscs of morals (1785)
- https://www.youtube.com/watch?app=desktop&v=nsgAsw4XGvU



Deontology

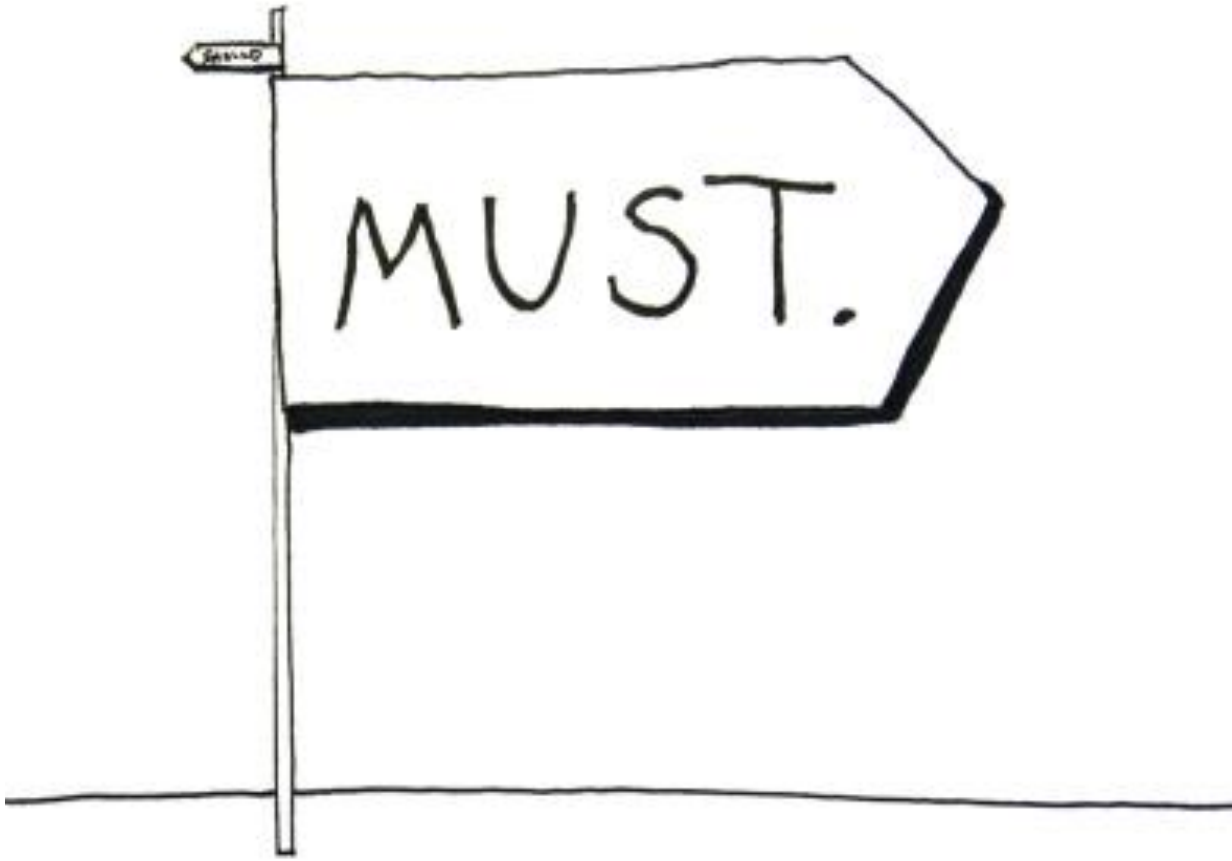
- Kantian ethics is a form of deontology or **deontology.**
- Deontology ➔ the normative ethical theory that the morality of an action should be based on whether that **action itself** is right or wrong under a series of rules and principles, rather than based on the consequences of the action.
- Duty ethics.



TILBURG UNIVERSITY

The Categorical Imperative (1)

- Categorical imperative ➔ the supreme principle of morality.
- Absolute moral Rules ➔ an action is morally right if it is in agreement with a rule that is applicable in itself.
- Deontology = it is not about the outcome, but about the intention/the rule underpinning the action



The Categorical Imperative (2)

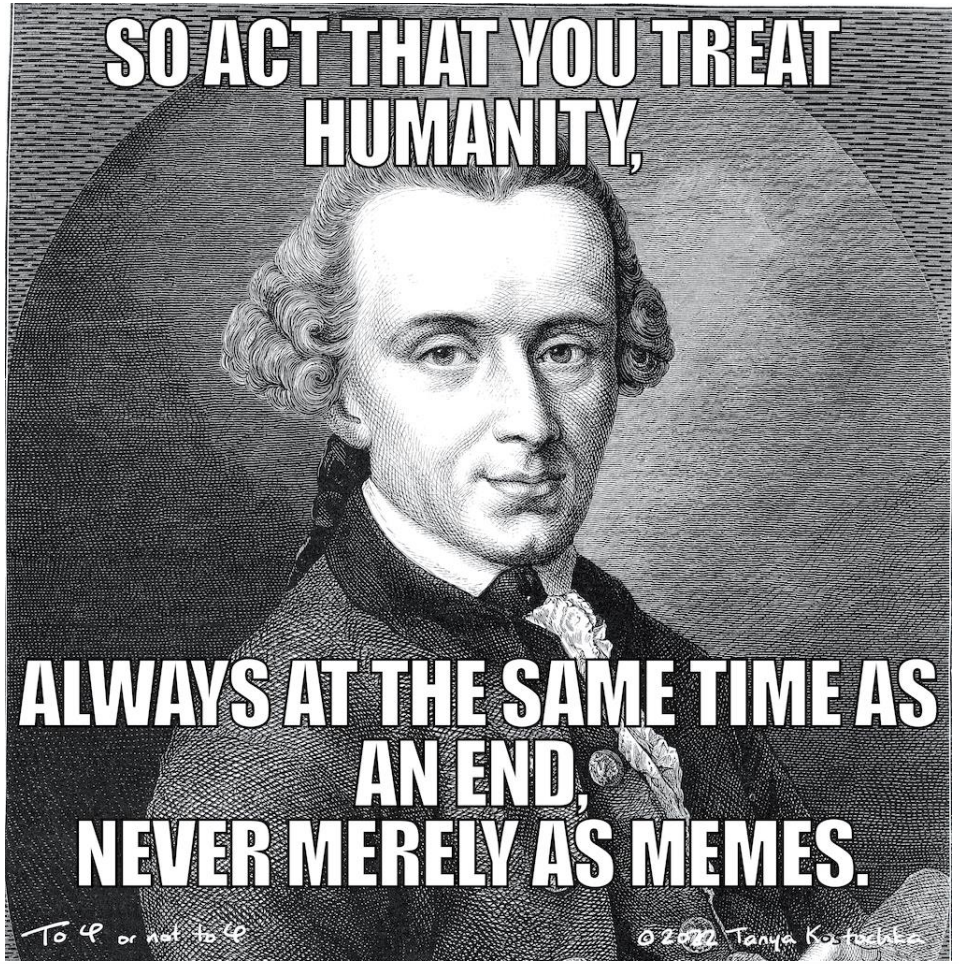
- It is categorical ➔ unconditional
  - You ought to do it, **no matter what**
  - It is a moral obligation (binding). In contrast to hypothetical imperatives (non-binding).
- It is an imperative
  - Human beings are not perfectly rational. They need to formulate a maxim based on which they can act.

The Categorical Imperative (2)

- Hypothetical “oughts” are possible because we have desires.
- Categorical “oughts” are possible because we have reason.
  - Derived from a principle that every rational agent needs to accept.
    - [compare to ‘public reason’]

The Categorical Imperative (3)

- “Act only according to that maxim by which you can at the same time will that it should become a universal law.” (universality principle)
- “So act that you use humanity, in your own person as well as in the person of any other, always at the same time as an end, never merely as a means.” (reciprocity principle)
- [“every rational being must so act as if he were through his maxim always a lawmaking member in the universal kingdom of ends.” (Autonomy)]

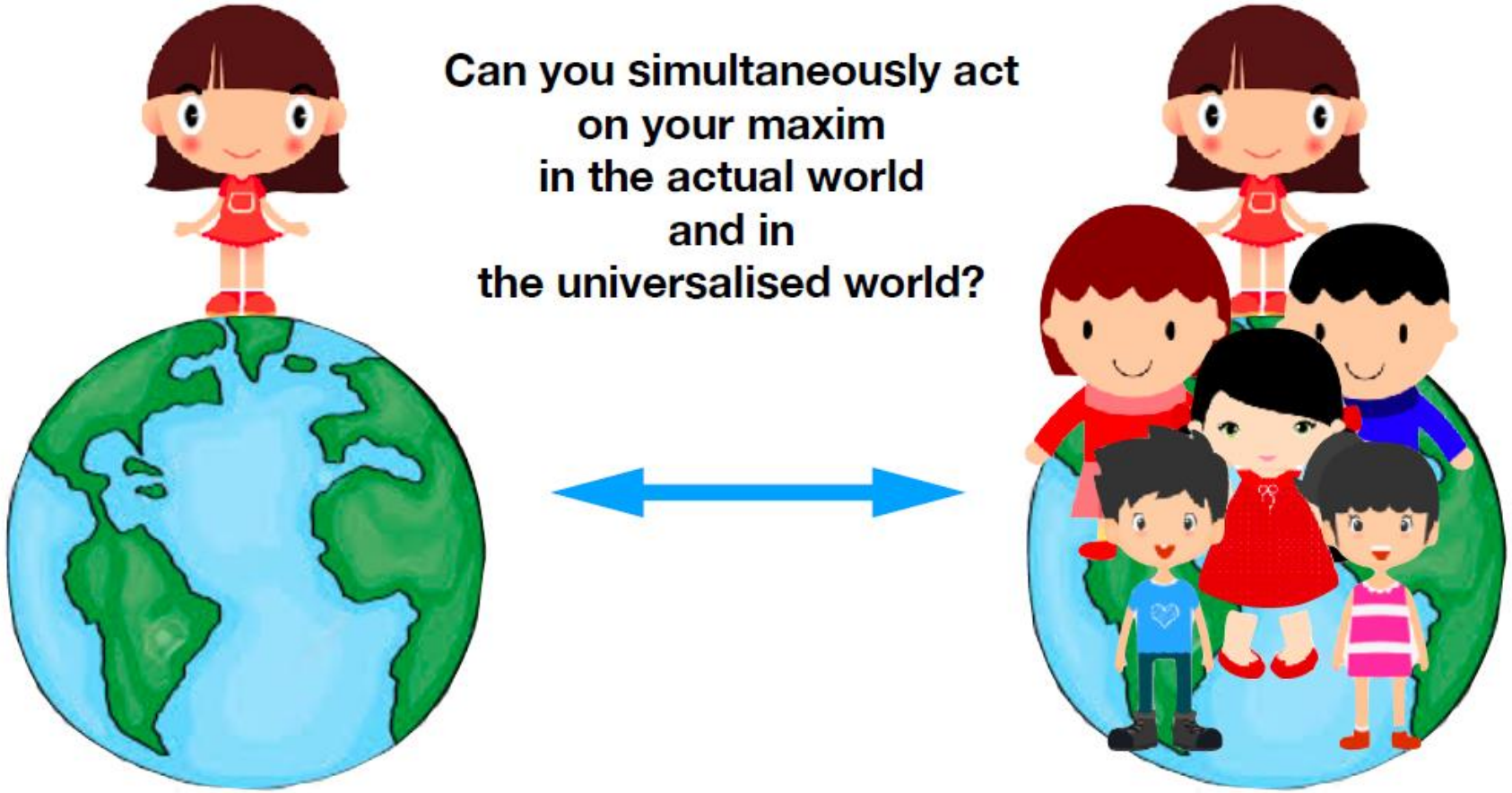


Step-by-step Procedure (CI 1)

How to act like a Kantian?

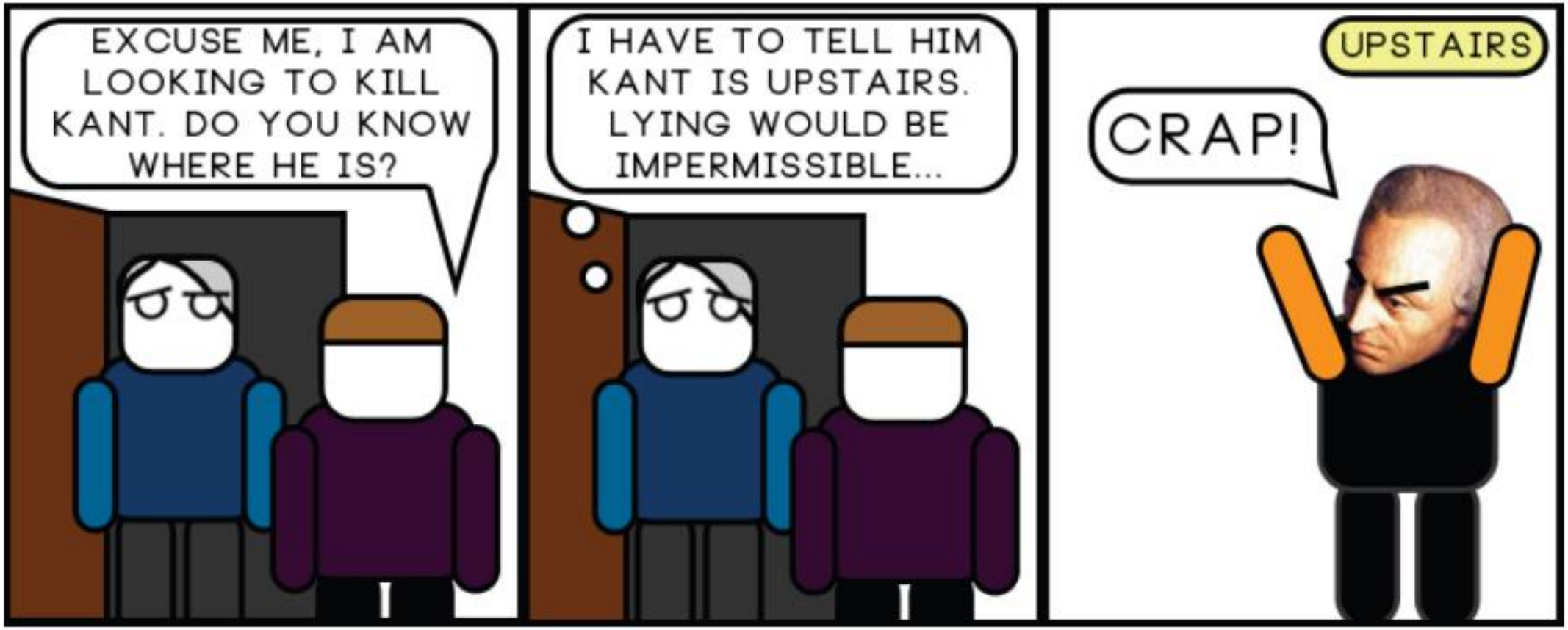
1. What rule are you following? Formulate the maxime.
2. Would you want that rule to be followed by everyone all the time? Formulate the universal law.
3. Does a contradiction arise?
   - A) No > go for it!
   - B) Yes > the action is impermissible

Categorical Imperative-test (CI 1)



The inquiring murderer

- A friend knocks at your door. She says she is chased by a murderer! You let her in so she can hide at your place. Next, the murderer knocks at your door and asks if your friend is at your place.
- What should you tell him, from a Kantian point of view?

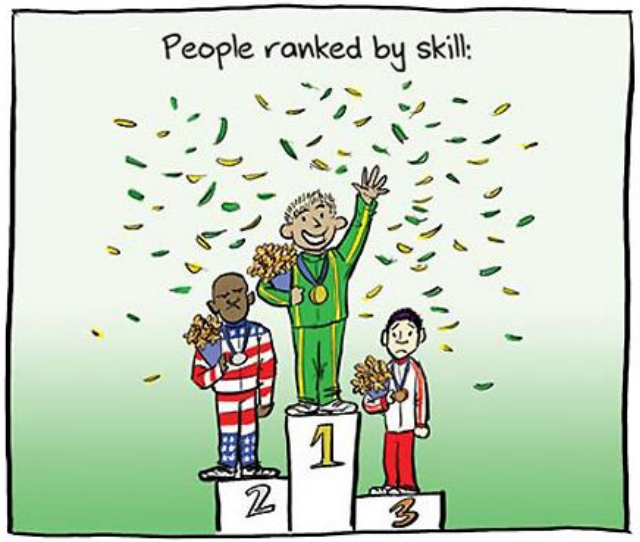

EXCUSE ME, I AM LOOKING TO KILL KANT. DO YOU KNOW WHERE HE IS?
I HAVE TO TELL HIM KANT IS UPSTAIRS. LYING WOULD BE IMPERMISSIBLE...
UPSTAIRS
CRAP!

Objections

- Are non-contradictory universal maxims possible?
  - William David Ross: ‘prima facie norms’ & ‘self-evident norms’
  - Hierarchy of norms
- Should consequences not be taken into account at all?
  - Example: child labor

Humanity as an end in itself (CI 2)

- “Act so that you treat humanity, whether in your own person or in that of another, always as an end and never as a means only”.



- Respect the rationality of other human beings.
- Human beings have an intrinsic worth, going above all price ➔ dignity.
- To use people is to disrespect their humanity.

Connection between morality and rationality

- CI is binding on all rational agents because they are rational.
- A moral judgment must be backed by good reasons.
- Acting immorally is acting irrationally.
- If you accept any considerations as reasons in one case, you must also accept them as reasons in other cases.
- If we would violate a rule, we do so for a reason that we would be willing for anyone to accept.

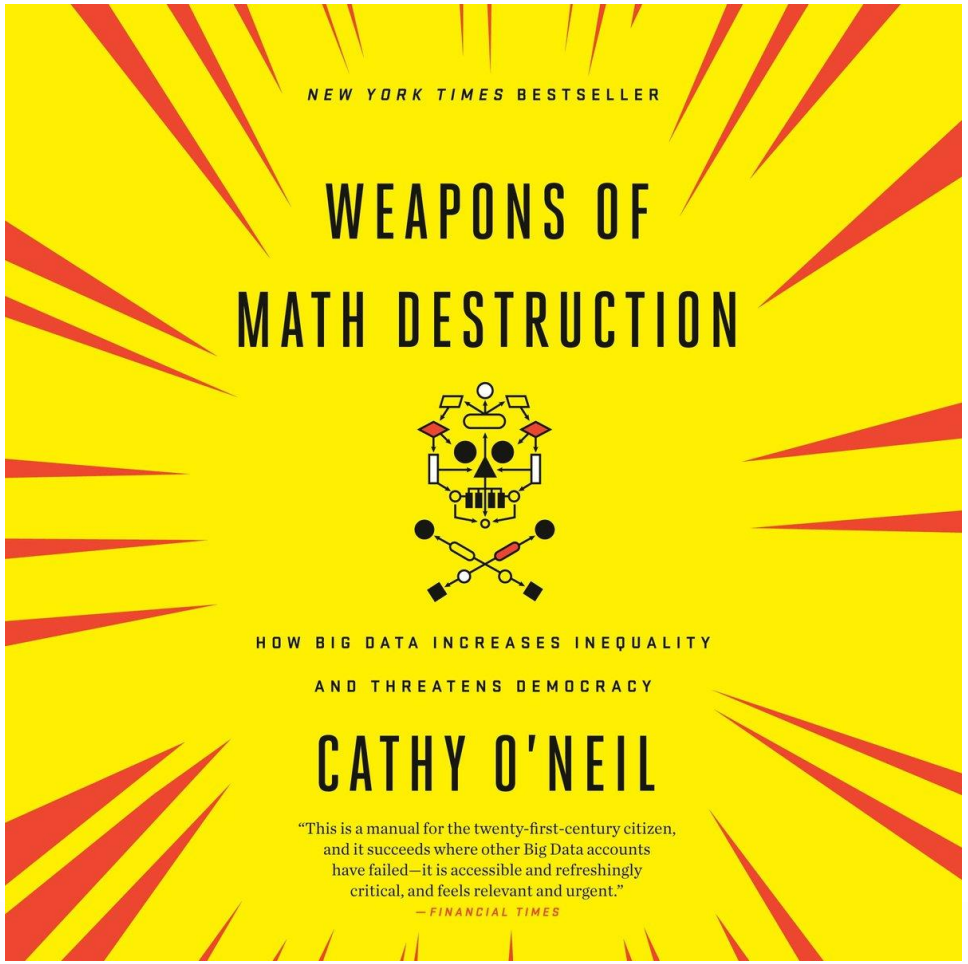
Risk profiling act ‘SyRI’ off the table

- Bring Human Rights Home: A Story from the Netherlands (youtube.com)
- SyRI: automated riskprofiling system that based on all kinds of different data sources scores people => worth investigating. (e.g. fraud)
- February 2020, the regional court in The Hague ruled that provisions for legal basis of SyRI are invalid.

- Child benefit scandal:
- Toeslagenaffaire: Cheikhou was jarenlang doelwit belastingdienst (youtube.com)
- Dutch scandal serves as a warning for Europe over risks of using algorithms – POLITICO

Cathy O'Neil (2016) Weapons of math descruction

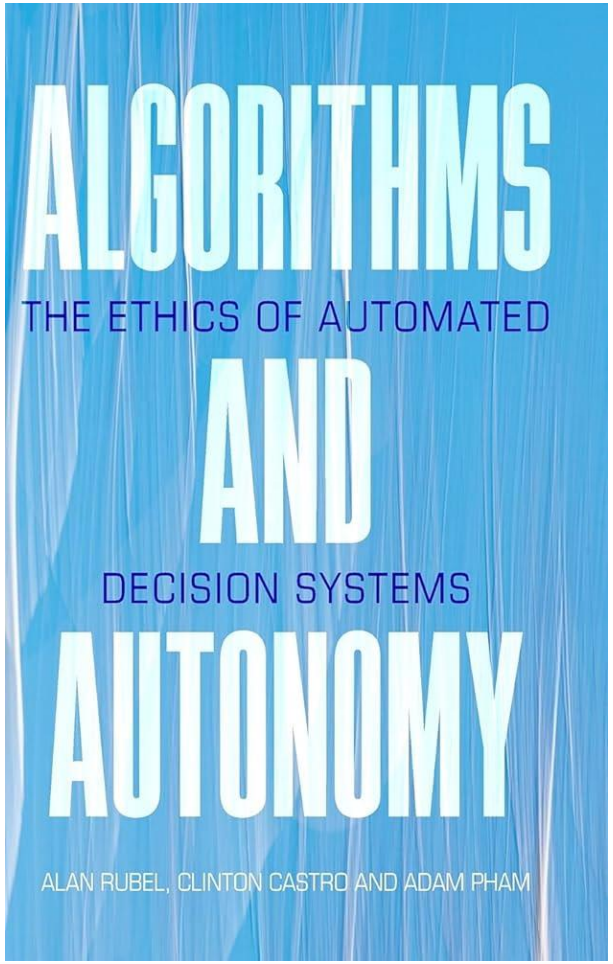
- Good models do exist
  - Transparant
  - Respond to feedback
- WMD's
  - Opaque
  - Beyond dispute or appeal
  - 'They do not listen. Nor do they bend. They're deaf not only to charm, threats, and cajoling but also to logic – even when there is good reason to question the data that feeds their conclusions.' (p. 10)



TILBURG UNIVERSITY

Automated Decision Making and autonomy

- Harm beyond negative consequences
  - Paying fine when committing fraud has negative consequences, but not a moral problem [?]
- Respecting persons, what we owe to them
- Ensuring the conditions of agency



Tilburg University

Algorithmic accountability

- Algorithmic accountability: providing reasons, explanations and justifications for decisions, including epistemic and normative assumptions.
  - Is it fair to treat a citizen in a certain way based on an aggregate analysis of behavioural data of other people?
  - On what ground do we ensure equality?
- Problem: Gap between what data scientists (or decision makers) see as an adequate justification and what those affected by the ADM-tool find acceptable.

Public reason as a constraint on ADM power

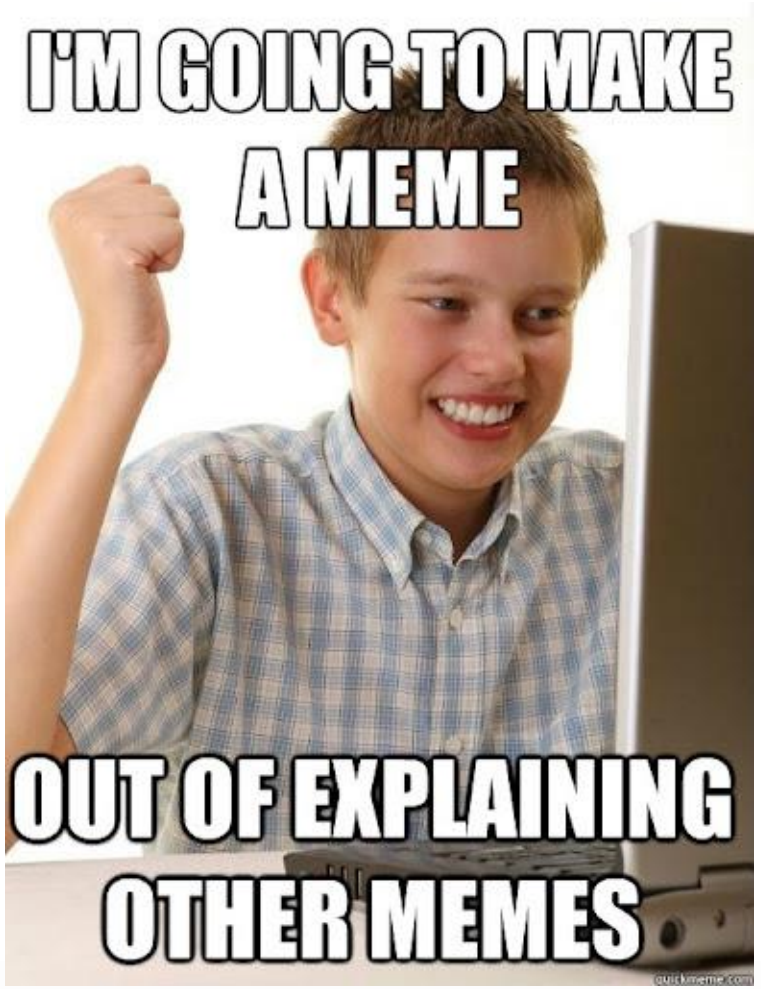
- Public reason: moral or political rules that regulate our common life should be, in some sense, justifiable or acceptable to all reasonable persons over whom the rules have authority.
  - Set of beliefs that are universal enough: e.g. equal basic liberty, equality of opportunity and a just distribution of income and wealth. (Binns 2018)
- account for the system's outputs according to normative standards which are acceptable to all reasonable people.

TILBURG UNIVERSITY

Public reason as a constraint on ADM power (2)

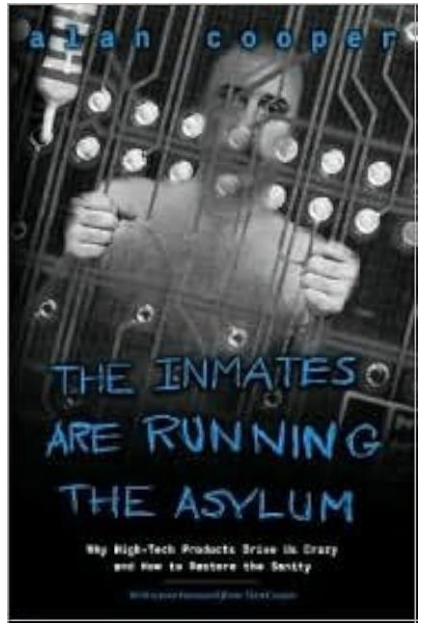
Advantages

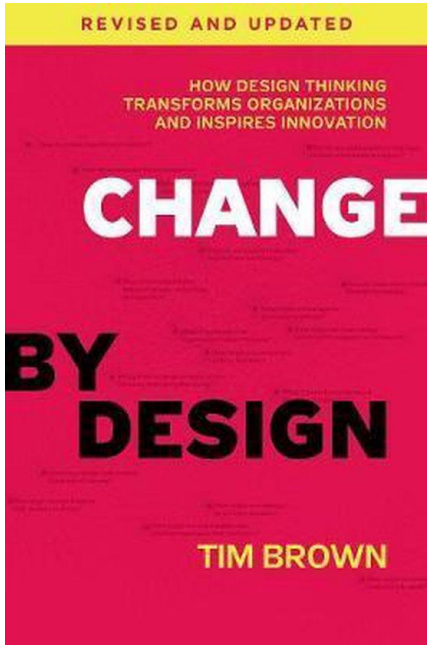
- Old, 'analogue' evaluations based on public reason (e.g. discrimination) also apply to ADM cases.
- It forces decision-makers to consider the ethical and epistemic aspects of their algorithmic systems ex ante
- Constrains both decision-makers as well as decision-subjects



What makes an explanation a meaningful explanation?

- Difference between explainability and explanation
  - Challenge for data scientists ("inmates running the asylum")
- What can we learn from social sciences, psychology and philosophy?



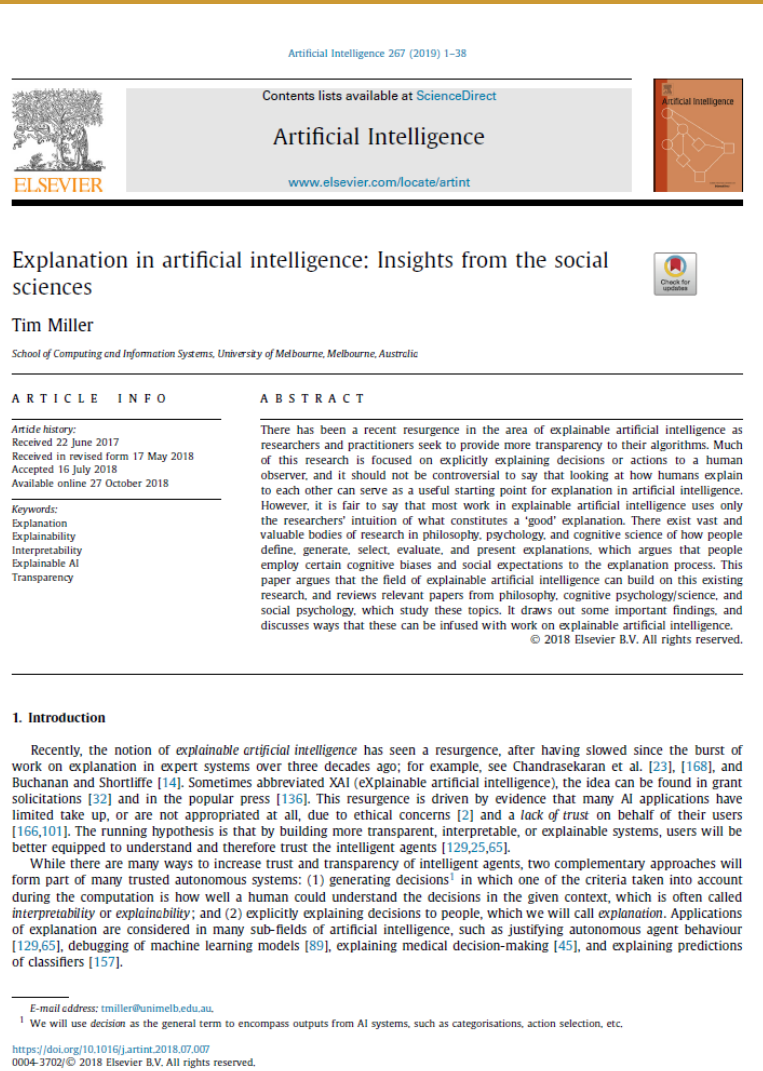


RTFM!

4 criteria

- Explanations are **contrastive**
- Explanations are **selected**
- **Probabilities**, probably, don't matter
- Explanations are **social**

➔ Explanations are contextual



Artificial Intelligence 267 (2019) 1–38

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Explanation in artificial intelligence: Insights from the social sciences

Tim Miller

School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

ARTICLE INFO

Article history:
Received 22 June 2017
Received in revised form 17 May 2018
Accepted 16 July 2018
Available online 27 October 2018

Keywords:
Explanation
Explainability
Interpretability
Explainable AI
Transparency

ABSTRACT

There has been a recent resurgence in the area of explainable artificial intelligence as researchers and practitioners seek to provide more transparency to their algorithms. Much of this research is focused on explicitly explaining decisions or actions to a human observer, and it should not be controversial to say that looking at how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence. However, it is fair to say that most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation. There exist vast and valuable bodies of research in philosophy, psychology, and cognitive science of how people define, generate, select, evaluate, and present explanations, which argues that people employ certain cognitive biases and social expectations to the explanation process. This paper argues that the field of explainable artificial intelligence can build on this existing research, and reviews relevant papers from philosophy, cognitive psychology/science, and social psychology, which study these topics. It draws out some important findings, and discusses ways that these can be infused with work on explainable artificial intelligence.

1. Introduction

Recently, the notion of *explainable artificial intelligence* has seen a resurgence, after having slowed since the burst of work on explanation in expert systems over three decades ago; for example, see Chandrasekaran et al. [23], [168], and Buchanan and Shortliffe [14]. Sometimes abbreviated XAI (eXplainable artificial intelligence), the idea can be found in grant solicitations [32] and in the popular press [136]. This resurgence is driven by evidence that many AI applications have limited take up, or are not appropriated at all, due to ethical concerns [2] and a *lack of trust* on behalf of their users [166,101]. The running hypothesis is that by building more transparent, interpretable, or explainable systems, users will be better equipped to understand and therefore trust the intelligent agents [129,25,65].

While there are many ways to increase trust and transparency of intelligent agents, two complementary approaches will form part of many trusted autonomous systems: (1) generating decisions[1] in which one of the criteria taken into account during the computation is how well a human could understand the decisions in the given context, which is often called *interpretability* or *explainability*; and (2) explicitly explaining decisions to people, which we will call *explanation*. Applications of explanation are considered in many sub-fields of artificial intelligence, such as justifying autonomous agent behaviour [129,65], debugging of machine learning models [89], explaining medical decision-making [45], and explaining predictions of classifiers [157].

E-mail address: tmiller@unimelb.edu.au.

[1] We will use *decision* as the general term to encompass outputs from AI systems, such as categorisations, action selection, etc.

https://doi.org/10.1016/j.artint.2018.07.007
0004-3702/

TILBURG UNIVERSITY

Next week: Virtue Ethics

- Ibo van der Poel & Lamber Royakkers (2011). Virtue Ethics in Ethics, Technology, and Engineering an introduction. https://cdn.prexams.com/6229/BOOK.pdfLinks to an external site.
- Ben Green (2021). Data Science as Political Action: Grounding Data Science in a Politics of Justice. https://scholar.harvard.edu/files/bgreen/files/data_science_as_political_action.pdf Links to an external site.

Questions on Module 3?

G.Meyers@tilburguniversity.edu