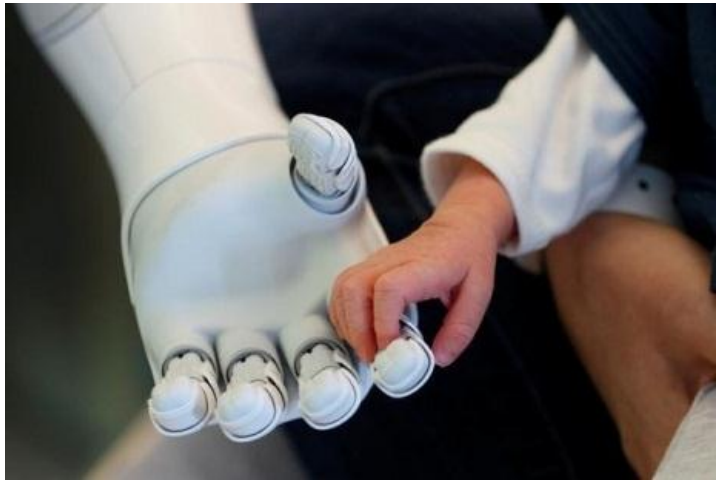
What's the hype about ethical AI?



Bettina Berendt

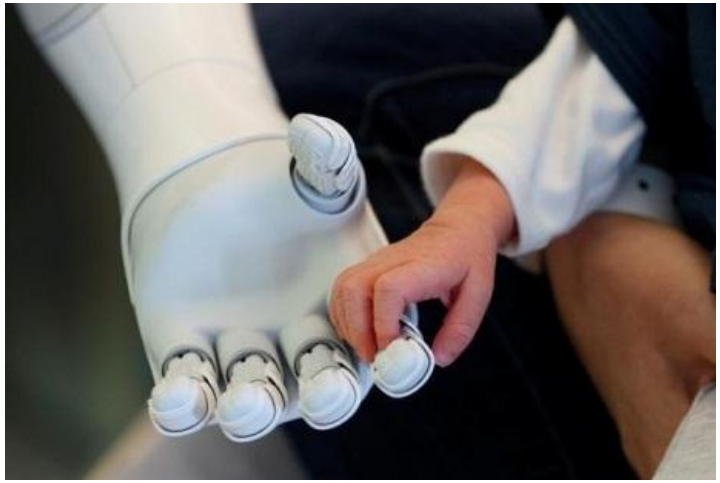
https://people.cs.kuleuven.be/~bettina.berendt

MLSS 2019

17 January 2019

What's the hype about ethical AI?

or: don't be cheesy



Bettina Berendt

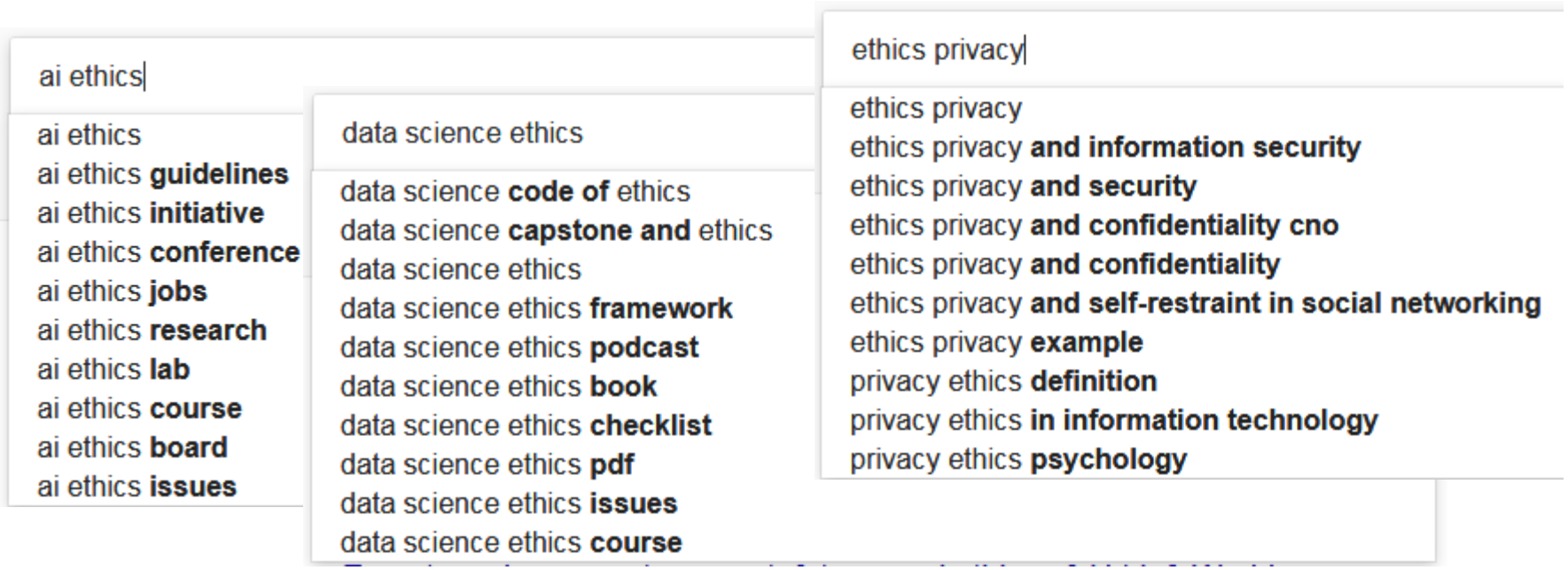
https://people.cs.kuleuven.be/~bettina.berendt

MLSS 2019

17 January 2019

Motivation (1):
Ask Google

“About 61.200.000 / 262.000.000 / 289.000.000 results”



“Consequently, the role of Ethics Advisors (EAs) and Ethics Advisory Boards (EABs) should be seen as the EC fulfilling its obligations to help avoid public uneasiness towards science and to mitigate concerns where they exist.”

http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/ethics-guide-advisors_en.pdf

Motivation (2):
Ask MLSS participants
(thanks to Jakob Zeitler for the survey!)

1. Themes I plan to talk about a lot

Scope of ethics, specific ethical judgements

Does the field of ethics provide any concrete, implementable guidelines applicable to non conscious machines/decision algorithms? can they be used to judge the behaviour of a black box without looking inside?

Is it any less ethical to take away someone's livelihood (by being replaced by an AI) than allowing that AI to act maliciously (in a non-physical harmful way) towards that person until they lose their job?

Common Good, democracy, power

How to democratize AI instead of accelerating inequality
How to prevent AI reinforcing current trends of rising inequality, and ensure benefits accrue more evenly to society as a whole.
How should we prepare our society for the AI revolution?
Making sure that decisions that are morally "off limits" for humans are only delegated to machines if it's undoubtedly in the public's interest.
To understand how democracy and AI can cooperate or coexist.
What about transparency, privacy, security, more power to AI systems?

What about the impact of AI on the future society? Is it worth it?

Research ethics / scientific clarity and honesty / scientific practice

correlation vs causality

Why and how do we get social scientists to the table?

Interdisciplinary research to insure ethical AI

Hype

Can there be "too much" ethics in AI?
Consider if the premise of the talk is true

2. Themes I plan to talk about only briefly
(I've said more about them in other talks & papers)

Rights, laws, regulation
Are people's rights always respected? Is there regulation?
Who is to what degree responsible for the actions of AI?

Will we be able to avoid the use of autonomous weapons?

What safeguards or oversight should be put in place on the algorithms and the people making them in order to avoid unethical applications of the technology?

How do we create a governing structure that can keep up with the developments in tech?
How will we ensure no blackboxes are used when it comes to machine learning so that we can ensure no discrimination occurs?
Is there a "GDPR" equivalent (or any acts) in the current AI framework/space?
Who's accountable if an AI system kills somebody? [BB: I think this means "responsible"]
Assuming that companies are generally held accountable for injuries caused by their products (e.g. f
Biases, (un)fairness, inequality

How should we curate representative datasets for entire populations, and how do we prevent biase

Avoid encoding human bias into the system. And where not possible, what biases are tolerable
Fairness ML decisions
The intersection should fairness and just.
Being aware and responsible when creating models relating to bias, impact on society. Being accou
Currently, fairness in ML
Having unbiased databases to train the algorithms
Can current and future AI ages guarantee equal opportunity, equal representation in perpetuity

No unfair biases
Ensuring automatic processing systems are fair with respect to sensitive attributes as well as ensuri

What are techniques to prevent algorithms to be "unfair" when there is an imbalance in the dataset

Can this discussion really start without challenging our understanding/perception of "cause and effe
biases and make efforts to address them in the algorithms. Those algorithms surely must include the
"cause and effect" (which is reality). How can we start without addressing this?
Counterfactual fairness

What is your opinion on Norman AI from MIT Media Lab? What can we learn from this?

Consent, privacy and data protection
do not be harmful to privacy of individual
Data privacy
Should data that is acquired illegally/ without consent be used for training models?
What about transparency, privacy, security, more power to AI systems?

Accountability

How can we make private companies [*BB: probably missing here:* accountable] for their algorithms?

Being aware and responsible when creating models relating to bias, impact on society. Being accountable.
How do we interpret the predictions? Predictions are merely a recommendation/option we still have the choice to use it or not. The responsibility is more on us when it comes to the decision-making and less on the algorithm that produced the option/s.

Who should be held accountable when a machine makes a mistake?
Who's accountable if an AI system kills somebody?

Rights, laws, regulation

Biases, (un)fairness, inequality

Consent, privacy and data protection

Accountability

3. Themes I will at most mention (and that I simply don't know enough about)

Human life

Dealing with human lives
What is the value placed on human life?

Labour

How are we going to offer re-training for people whos wo
Something about job replacement!
The loss of jobs due to AI
What do you think the future of work looks like?

Machine ethics and related questions

What are the ramifications of AI being used for moral dec
Can AI be used to make ethical decisions on its own?
Who is to what degree responsible for the actions of AI?
Making the distinction can be hard for humans so how ca
How do we deal with self conscious AI?
Should intelligent systems be treated the same as humans
Is it ethical to expect an AI to conform to our norms and e
that reinforces ethical action in humans?
What rights does an AI system have.

Technical questions about AI, needed to ask further ques

How far has the development of autonomous weapons in
What is the current state on the research in the feild

Things I don't believe in

How should we handle AI that will supersede human intell

Questions that I did not understand

When a self-driving car will crash.

al traits

ork?

Human life

Labour

Machine ethics and related questions

Technical questions about AI, needed to ask further questions about that AI's ethical dimension

Things I don't believe in

Questions that I did not understand

Positioning, caveat, and invitation

- My case studies and thinking are German-, EU- and a bit US-centric …
- This ignores:
  - "AI for Africa" (M. Hoffmann, ca. 2014)
  - the wide range of backgrounds in MLSS
- I'd love to hear about your case studies, laws, …

Part 1:
Ethics is about doing “good”,
or is it?

A data-science case study

- Havasupai: a small, isolated tribe in AZ
- 1990: blood samples for diabetes research
- Consent
- 2003:
  - a tribe member learns about re-use for theories about the tribe's ancestors' geographical origins, kinship structures, etc.
  - "*deeply offensive*"
  - "banishment order" for U.Arizona
- 2010: after $1,700,000 in legal fees, the university agreed to pay $700,000, and the samples were destroyed.





The ethical scientist?

- The Havasupai reclaimed the 151 remaining blood samples.
- The geneticist:
  - *"I was doing good science."*
  - Claim: Those judging her [as non-ethical] fail to understand the fundamental nature of genetic research, where progress often occurs from studies that do not appear to bear directly on a particular disease.
- *It has since been claimed that scientific study of Native Northern American genomes has been hampered due to scientists' fears of getting caught up in a similar legal predicament.*





Take-home message #1

- Acting ethically means doing “good”.
- But “good” is not a well-defined concept.
- You should identify stakeholders and learn about their notions of “good” and “bad”.
- Be prepared to learn. And to re-learn.
- Be prepared to make mistakes. And to deal with them.

If your data science project …

… may help cure diabetes for 800 people

… but may reveal inbreeding, considered a deadly sin, in some families (40 people)

Would you do it?

If your data science project …

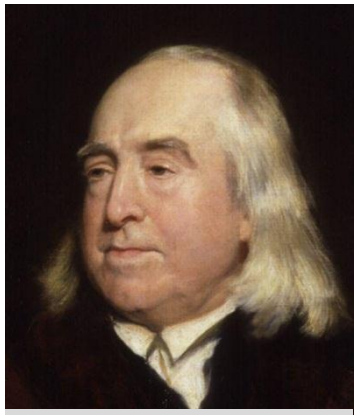
… may help cure diabetes for 800 people

… but may reveal your best friend's father is not their biological father

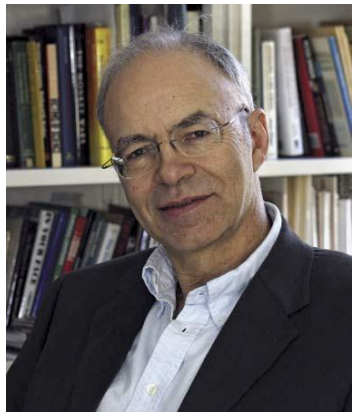
Would you do it?

What do you maximise?

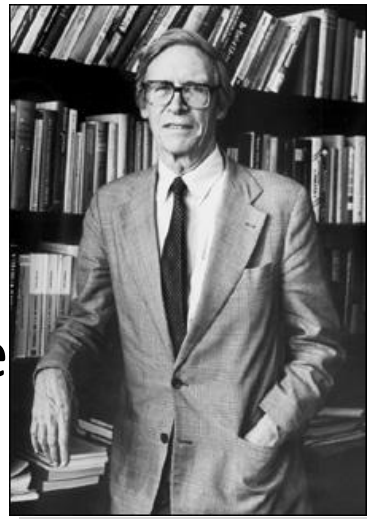
- And what determines the overall "goodness" (utility) of your project?
- *Average*: The sum, over all stakeholders, of the happinesses and unhappinesses?
- *Minimum*: The lot of the stakeholder who's worst off?
- *Moral universe; justice or care?*: Over which stakeholders to aggregate? Do some "weigh" more and why?



Jeremy Bentham



Peter Singer



John Rawls



Nel Noddings

If your data science project …

… may help/save the lives of 1000s of people

… but will lead to the death of 10 people

Would you do it?

Take-home message #2



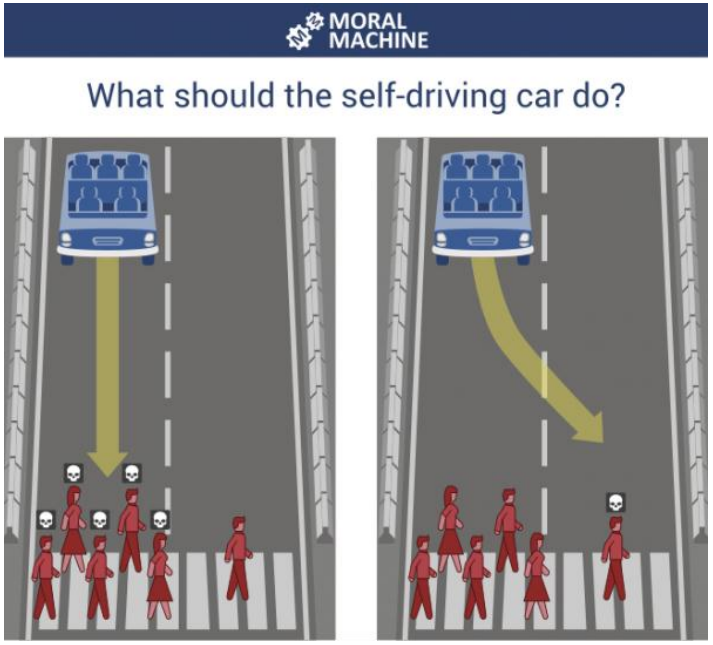
Immanuel Kant

- Aggregation over stakeholders can take different forms.
  (assuming utilitarian theories are used)
- Fundamental principles / rights, in particular human dignity, may forbid aggregation/trade-offs.
- Stakeholder-centrism and aggregation may not be applicable – some things are "simply wrong".
  (≠ Descriptive ethics à la Awad et al., *Nature* 2018)





Philippa Foot



If you work in a data-science project shaped by your boss/supervisor/etc.

- And you conclude that this project violates ethical standards (and/or legal ones)
- But saying this in public also violates laws, and may endanger your job, comfort, …
- … would you:

- Talk to your data protection officer?
- Go public?
- Not do anything?

What if the DPO does nothing?

What if the DPO does nothing?

I believe in the principle declared at Nuremberg in 1945: "Individuals have international duties which transcend the national obligations of obedience. Therefore individual citizens have the duty to violate domestic laws to prevent crimes against peace and humanity from occurring."



Edward Snowden

Take-home message #3

- Ethical thinking cannot be fully delegated.
- *You can always do something.*

- *We must use critical thought to distinguish what is ethical from what is lawful and to consider what it means to be a professional. Therefore, we must continually question and re-question authority, whether it is the law or a code of ethics, or else we may be doomed to serve the interests of those who crafted the code […]. Just because a principle is codified does not make it ethical.*

  (Patel & Elkin, 2015, on the APA Ethics Code / CIA torture case)



Hannah Arendt

Summary: Different approaches to ethical judgement

- Consequentialist (often: utilitarian) ethics
- Deontological ethics
- Ethics of Care
- …

Example: how does this help us answer the question / reason about the question (from the MLSS survey)

- *Is it any less ethical to take away someone's livelihood (by being replaced by an AI) than allowing that AI to act maliciously (in a non-physical harmful way) towards that person until they lose their job?*

Good for all?!

“Alignment with human values” or “good for some” – not good enough!

Even lethal autonomous weapons are “good” for some!

➔ Typical AI Ethics requirement: *Common* Good (CG)



- Examples from the Asilomar Principles:
  - Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.
  - Shared Benefit: AI technologies should benefit and empower as many people as possible.
  - Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
- Similar in Ethics Codes for Comp.Sci., AI by ACM, IEEE, …

(Berendt, 2019)

How >2000 years of philosophy can help us question Ethical AI (yet again)

The Common Good (CG): that which benefits society as a whole

"*that*": The relevant interests and facilities that serve these interests together constitute the CG and serve as a shared standpoint for political deliberation [among the members of a political community].
These facilities include social institutions, rights, human rights.

(Berendt, 2019)

How >2000 years of philosophy can help us question Ethical AI (yet again)

The Common Good (CG): that which benefits society as a whole

"*that*": The relevant interests and facilities that serve these interests together constitute the CG and serve as a shared standpoint for political deliberation [among the members of a political community].
These facilities include social institutions, rights, human rights.

Dimensions of / challenges for CG	
Benefit: "utility" or values beyond this? Who is (not) in the *community/society*?	
Substantive vs. procedural	
Communal vs. distributive	
Interests and facilities	

(Berendt, 2019)

How >2000 years of philosophy can help us question Ethical AI (yet again)

The Common Good (CG): that which benefits society as a whole

"*that*": The relevant interests and facilities that serve these interests together constitute the CG and serve as a shared standpoint for political deliberation [among the members of a political community].
These facilities include social institutions, rights, human rights.

Dimensions of / challenges for CG	Lead questions for AI for the CG
Benefit: "utility" or values beyond this? Who is (not) in the *community/society*?	Q1: What is the problem?
Substantive vs. procedural	Q2: Who defines the problem?
Communal vs. distributive	Q4: What are important side effects and dynamics?
Interests and facilities	Q3: AI as engineering → a facility → What is the role of knowledge?

(Berendt, 2019)

Reality check: Aren't all these Qs obvious and dealing with them part of standard practices?

- Exploratory study of all 99 hyperlinked contributions to 4 current major venues on AI for Good, Data Science for Social Good, …

- Definition of "the Good": mainly substantive
- Q1: Alternative goals?
- Q2: Different stakeholders, description of how their perspectives and needs were assessed?
- Q3: The role of knowledge?
- Q4: Dynamics?

(Berendt, 2019)

Reality check: Aren't all these Qs obvious and dealing with them part of standard practices?

- Exploratory study of all 99 hyperlinked contributions to 4 current major venues on AI for Good, Data Science for Social Good, …

- Definition of "the Good": mainly substantive
- Q1: Alternative goals?
- Q2: Different stakeholders, description of how their perspectives and needs were assessed?
- Q3: The role of knowledge?
- Q4: Dynamics?

} ✘ or in very few con-tribu-tions

(Berendt, 2019)

NB: I believe that some of this neglect is a consequence of our project & publishing culture

- Exploratory study of all 99 hyperlinked contributions to 4 current major venues on AI for Good, Data Science for Social Good, …

- Definition of "the Good": mainly substantive
- Q1: Alternative goals?
- Q2: Different stakeholders, description of how their perspectives and needs were assessed?
- Q3: The role of knowledge?
- Q4: Dynamics?

} ✘ or in very few con-tribu-tions

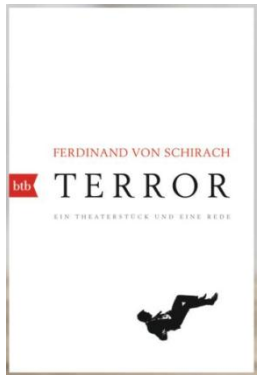
(Berendt, 2019)

Some thoughts on “militant democracy”, or: Why don’t we just crowdsource The Universal Declaration of Human Rights 2.0?

By this stage, you should be able to reason about the hidden assumptions in this conclusion of the Awad et al. Moral Machine experiment paper (p. 63):

“Indeed, we can embrace the challenges of machine ethics as a unique opportunity to decide, as a community, what we believe to be right or wrong; and to make sure that machines, unlike humans, unerringly follow these moral preferences.

We might not reach universal agreement: even the strongest preferences expressed through the Moral Machine showed substantial cultural variations, and our project builds on a long tradition of investigating cultural variations in ethical judgments.”



Some thoughts on bias, unfairness, and discrimination; or: "unfair or unjust?"

Ad related to **latanya sweeney** ⓘ

Latanya Sweeney Truth
www.instantcheckmate.com/
Looking for **Latanya Sweeney**? Check **Latanya Sweeney's** Arrests.



Ads by Google

Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
Public Records Found For: **Latanya Sweeney**. View Now.
www.publicrecords.com/

La Tanya
Search for La Tanya Look Up Fast Results now!
www.ask.com/La+Tanya

(Sweeney, 2013), more in (Berendt & Preibusch, 2014, 2017) and related talks on my homepage

So do we have any “solution”?

So do we have any “solution”?

- Is this the right question?
- Maybe *this* question prevents us from asking the other ones?!

Problems and problem solving

	Engineering **“Problem solving approach”**
Goal function	well-defined
Method	can be black box
Decomposition, Delegation	modular; can be fully delegated
Ethics codes and assessments	Canonical; often a burdensome afterthought
Solvability	Problems can be solved.

(Berendt, 2019)

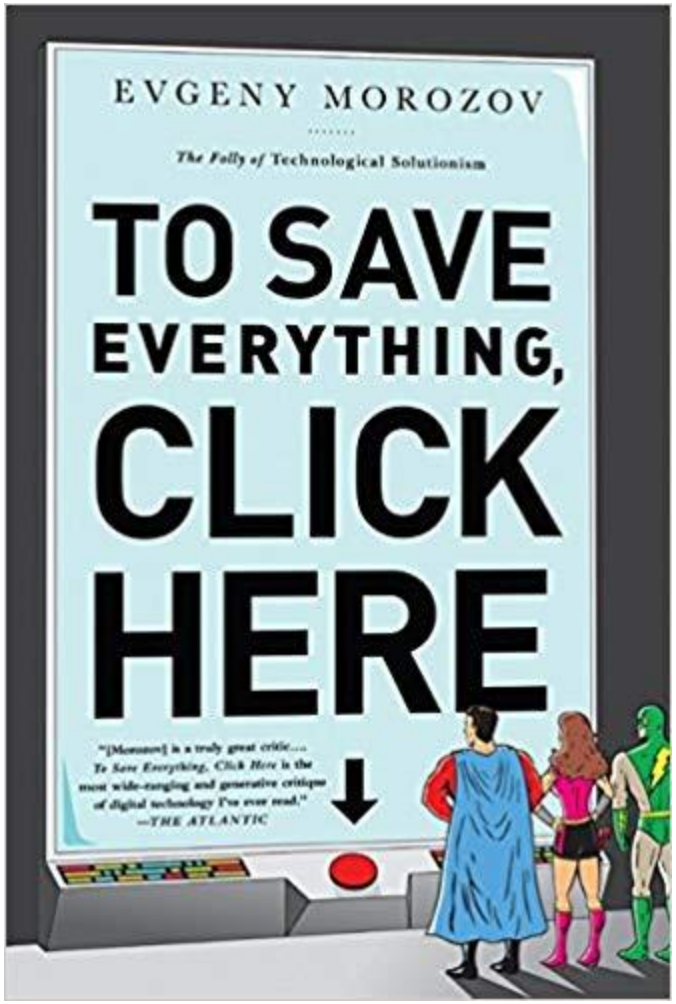
Social problems are not chess problems

	Engineering "Problem solving approach"	Dealing with social problems
Goal function	well-defined	often involves unresolvable ethical dilemmas, continued re-negotiation
Method	can be black box	requires fairness, transparency, accountability
Decomposition, Delegation	modular; can be fully delegated	Usually interdependent, (at least some) participation required
Ethics codes and assessments	Canonical; often a burdensome afterthought	starting points; "ethics by design"
Solvability	Problems can be solved.	Some problems can only be addressed.

"Solutionism"

(Berendt, 2019)

So do we have any “solution”?



… or at least “approaches”?

Each DMP needs to adhere to a number of requirements that are outlined below. Our DMP builds further on the KU Leuven standardized template. Use the DMPonline.kuleuven.be service to create your DMP. The DMPs in the online tool are only available to yourself and those people you add to your DMP (e.g. your supervisor).

Especially in the case of personal data, KU Leuven requires you to register the dataset and the research lab requires a full DMP such that there is a complete record if there would be an audit. You need to be able to show that you have performed a Data Protection Impact assessment and have mitigated potential risks.

3.1 Administrative Data

Here you should record basic information to identify and contextualise your plan. Identifiers may help to link your DMP with information held in other systems. You should include:

- Basic information e.g. project title, your name, contact details, reference numbers / IDs
- A summary of the research to explain the **purpose** and **scope** for which data are being collected.
- Details of related policies and procedures e.g. institutional data policy or departmental guidelines.

3.2 Data Collection

What data will you collect or create?

Describe clearly:

- The type of the data and the categorisation (Personal/Non-Disclosure/Proprietary/Public)
- The format of the data
  - Preferably you use a known standard (e.g. csv, arff)
  - Do your chosen data formats and scripts enable sharing and long-term access to the data?
  - Refer to and see also Section 3.7 "Data Sharing".
- The size
- How it was generated

Do you intend to reuse existing data?

Describe clearly:

- The type of the data and the categorisation (Personal/Non-Disclosure/Proprietary/Public)
- The license of the data (this is mandatory!)
- The source/authors of the data
- In case of personal data, is your research topic considered primary or secondary use? Is secondary use allowed (e.g. mentioned in consent)?

How will you transfer third-party data to KU Leuven?

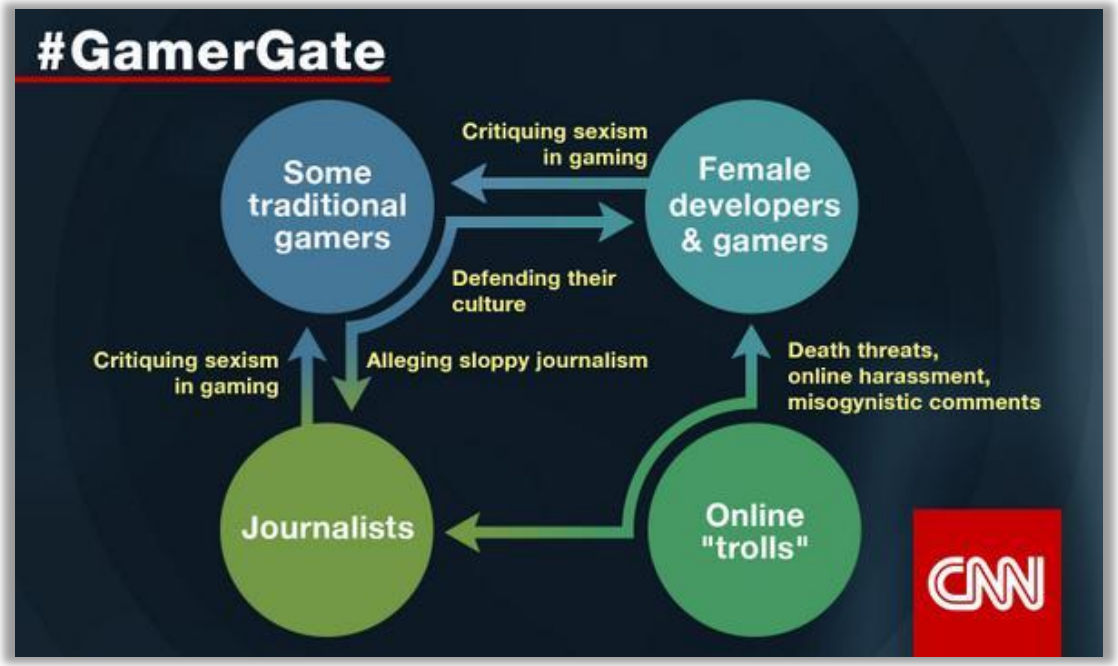
A1: Data governance with a focus on legal compliance

(De Raedt & Meert, 2017f.)

If your data science project …

… is about Gamergate

Would you do it?



Even Doing Academic Research On Video Games Puts Me At Risk

By A.D. Andrew

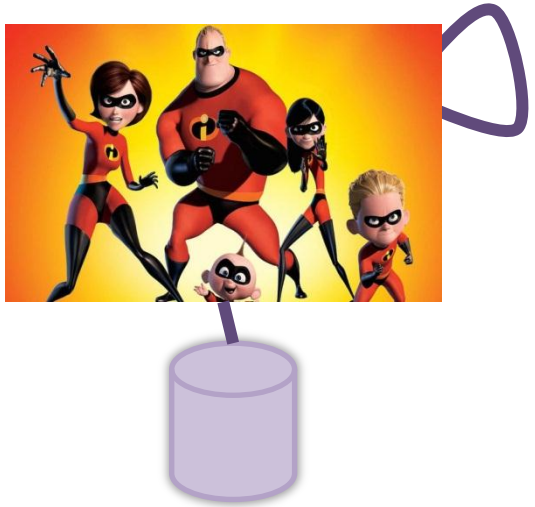
A2: An Ethics of Care (EoC) framework focused on research relationships and decision making in project design

- EoC 'vs.' justice-oriented approaches to ethics
  - EoC guides us in situations where rights don't apply or aren't clear
  - respect as precondition for rights
- Key elements
  - Relationships with power differentials rather than independent rational individuals with rights (focus on fairness and equality),
  - dialogue
  - practice rather than rules

This slide and the following ones: (Suomela, Chee, Berendt, & Rockwell, in press)

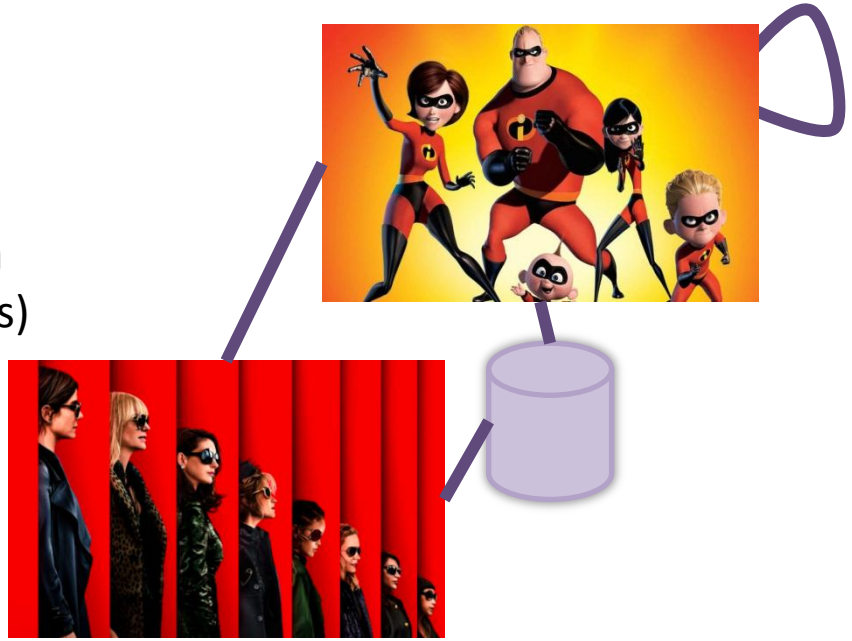
Applying an Ethics of Care to the case study: Relationships, dialogues, decisions

- Different vulnerabilities of / care for **team members**
  - External harms
  - Psychological harms
- Decisions about authorship attribution



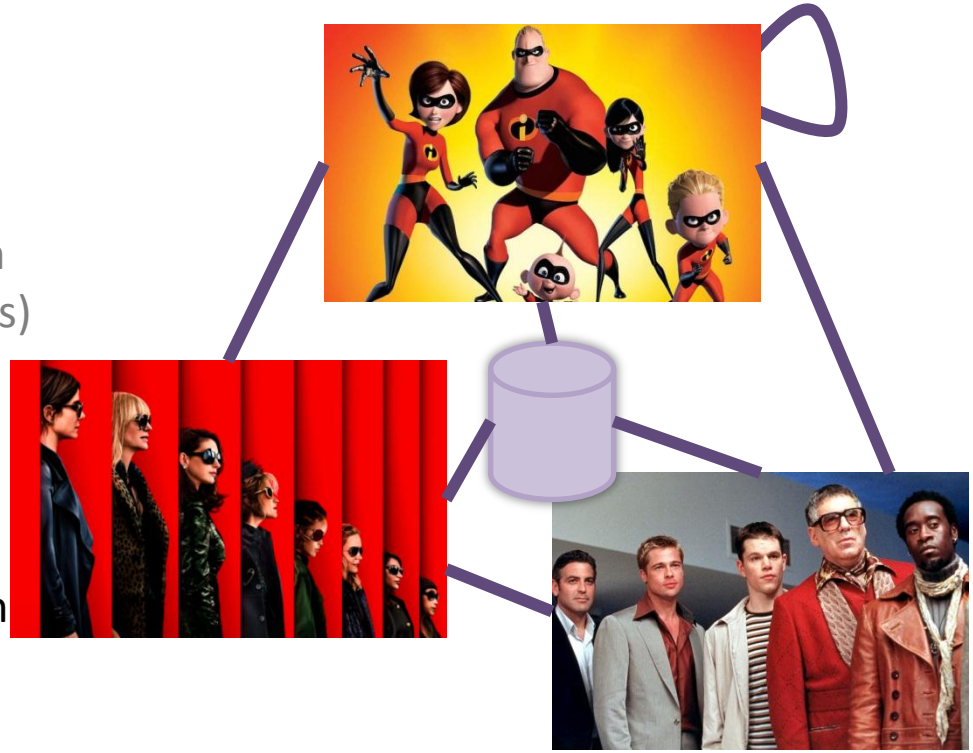
Applying an Ethics of Care to the case study: Relationships, dialogues, decisions

- Different vulnerabilities of / care for **team members**
  - External harms
  - Psychological harms
- Decisions about authorship attribution
- Toxic data: decision to not use (analysis) and delete (archive)
  - Harms to team members, **talked-about people**



Applying an Ethics of Care to the case study: Relationships, dialogues, decisions

- Different vulnerabilities of / care for **team members**
  - External harms
  - Psychological harms
- Decisions about authorship attribution
- Toxic data: decision to not use (analysis) and delete (archive)
  - Harms to team members, **talked-about people**
- **Authors**:
  - Privacy harms, reputation harms
- Attribute filtering and de-identification



Applying an Ethics of Care to the case study: Relationships, dialogues, decisions

- Different vulnerabilities of / care for **team members**
  - External harms
  - Psychological harms
- Decisions about authorship attribution
- Toxic data: decision to not use (analysis) and delete (archive)
  - Harms to team members, **talked-about people**
- **Authors**:
  - Privacy harms, reputation harms
- Attribute filtering and de-identification
- **Scientists / possible re-users of the data**
- Relationships vs. abstract duty towards science



Applying an Ethics of Care to the case study: Relationships, dialogues, decisions

- Different vulnerabilities of / care for **team members**
  - External harms
  - Psychological harms
- Decisions about authorship attribution
- Toxic data: decision to not use (analysis) and delete (archive)
  - Harms to team members, **talked-about people**
- **Authors**:
  - Privacy harms, reputation harms
- Attribute filtering and de-identification
- **Scientists / possible re-users of the data**
- Relationships vs. abstract duty towards science
- See results at https://dataverse.library.ualberta.ca/dataset.xhtml?persistentId=doi:10.7939/DVN/10253



Is this a solution?



"Finding a balance between … is an ongoing dilemma. … In the case of …, we decided ... Other researchers may have acted differently but we can only report the results of our own analysis and reflection upon the methods which produced that analysis."

(Suomela, Chee, Berendt, & Rockwell, in press)

Part 2:
Hype?

Hype

English
Oxford *Living* Dictionaries

- [noun] Extravagant or intensive publicity or promotion.
- [noun] A deception carried out for the sake of publicity.
- [verb] Promote or publicize (a product or idea) intensively, often exaggerating its benefits.

It is *good* that scientists (and practitioners, and …) think about ethics!



10 things that made 2018 great for the movement to end nuclear weapons

December 28th 2018

How do you top winning the Nobel Pea
and booking win after win. So

The **Declaration of Helsinki** (**DoH**, Finnish: *Helsingin julistus*, Swedish: *Helsingforsdeklarationen*) is a set of ethical principles regarding human experimentation developed for the medical community by the World Medical Association (WMA).[1] It is widely regarded as the cornerstone document on human research ethics.[1][2][3][4]

It is not a legally binding instrument under the international law, but instead draws its authority from the degree to which it has been codified in, or influenced, national or regional legislation and regulations.[5] Its role was described by a Brazilian forum in 2000 in these words "*Even though the Declaration of Helsinki is the responsibility of the World Medical Association, the document should be considered the property of all humanity*".[5]

The Biological Weapons Convention (BWC), the first multilateral disarmament treaty banning the development, production and stockpiling of an entire category of weapons of mass destruction, was opened for signature on 10 April 1972. The BWC entered into force on 26 March 1975.

Of course there's hype, because the context is hyped too

KDnuggets™

SOFTWARE | News/B

Over the past deca
seen striking developments. As surveyed in, there now exist over twenty domains in which AI programs are performing at least as well as (if not better than) humans.

The Current Hype Cycle in Artificial Intelligence

By Alok Aggarwal, Scry Analytics.

comments

Prologue

Every decade seems to have its technological buzzwords: we had personal computers in 1980s; Internet and worldwide web in 1990s; smart phones and social media in 2000s; and Artificial Intelligence (AI) and Machine Learning in this decade. However, the field of AI is 67 years old and this is the third in a series of five articles wherein:

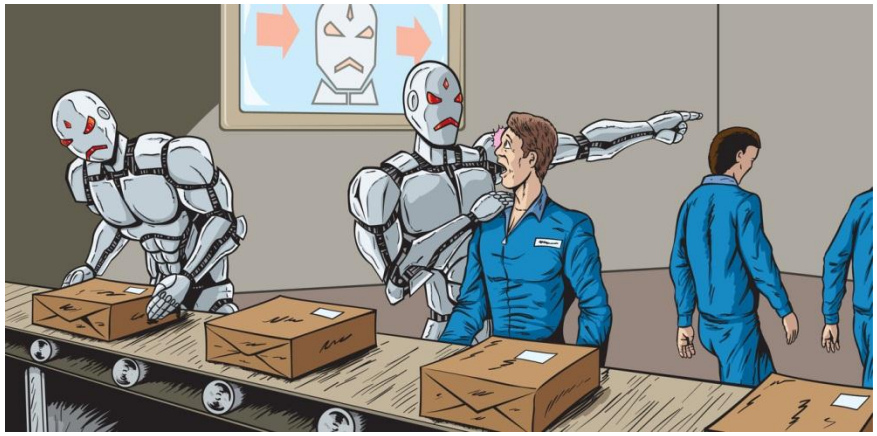
1. The first article discusses the genesis of AI and the first hype cycle during 1950 and 1982
2. The second article discusses a resurgence of AI and its achievements during 1983-2010
3. The third discusses the domains in which AI systems are already rivaling Humans
4. This article discusses the current hype cycle in Artificial Intelligence
5. The fifth article discusses as to what 2018-2035 may portend for brains, minds and machines

The Timeline



Risk: Focus on fears and cool AI, neglect of statistically larger risks

Example labour market



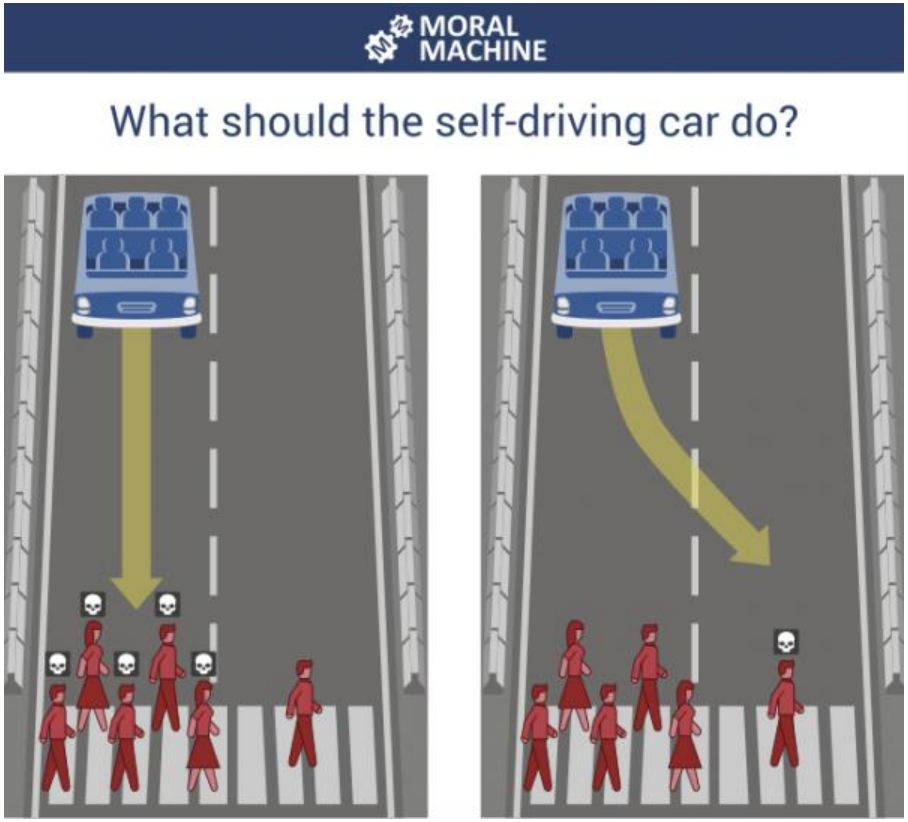
Fear:

- AI robots taking our jobs?

Facts:

- Relatively dumb algorithms and robots can have large impacts, e.g.:
  - Scheduling software → "clopening" shifts
  - Crowdsourcing → working standards, pay, …
  - Rating economy → emotional labour
  - US: 61 robot-related deaths, 1992-2015

Risk: Make assumptions that can become a self-fulfilling prophecy, neglect existing approaches to known problems



- Yes, more info on people available and processable at great speed can change environments, situations, decisions.
- A known problem!
- A known approach (e.g. in criminal law): inadmissibility of evidence

Risk: Furtive undermining of due process and the role of the state

Recap: If your data science project …

… may help/save the lives of 1000s of p

… but will lead to the death of 10 peop

Doesn't need to be life-or-death. This is a well-known problem for example regarding proprietary software for risk assessment in predictive policing.

Who makes, and
who is entitled to make,
such decisions?

The army?
The police?
A software company?
"An AI"?



Risk: fuzzy warm feelings rather than enforceable rights

"In the AI area … there's a tendency to talk about ethics which is undefined and unaccountable, conveniently."

(Philip Alston, AI Now Symposium, 2018)

"Ethics" (in its fuzzy sense)	Law, regulation, (human/fundamental) rights	But ethics is more than fuzziness – e.g.:
The magnanimity of the one who does good (scientists, developers, …)	The rights of the one who is entitled to the good (citizens, people, …)	(Human) rights are part of the Common Good, increasing attention (privacy & FAT research, FRA, …)
Industry self-regulation	Law, regulation, law enforcement – in a political system with checks & balances	Ethics and the law are neither identical nor disjoint, & complementary!

Some more thoughts about existing structured approaches: law and software engineering

Accountability: political science

(Bovens, 2007)

A relationship qualifies as a case of accountability when:

1. There is a relationship between an actor and a forum
2. in which the actor is obliged
3. to explain and justify
4. his conduct,
5. the forum can pose questions,
6. pass judgement,
7. and the actor may face consequences.

Privacy Eng./GDPR approach:
Accountability: a political science concept & the GDPR

A relationship qualifies as a case of accountability when:

1. There is a relationship between an actor and a forum
2. in which the actor is obliged
3. to explain and justify
4. his conduct,
5. the forum can pose questions,
6. pass judgement,
7. and the actor may face consequences.

"Technical and organisational measures" relating to

processing of personal data per se

GDPR obligations

Rec. 71, Art. 13-15, 20

Art. 5 (2)

Art. 13-15, 20, 22; 5(2)

) Art.

) 83, 58

AI/Machine Learning meets software engineering meets legal principles

Examples: Privacy design strategies and patterns and the VeriLearn project @ KU Leuven

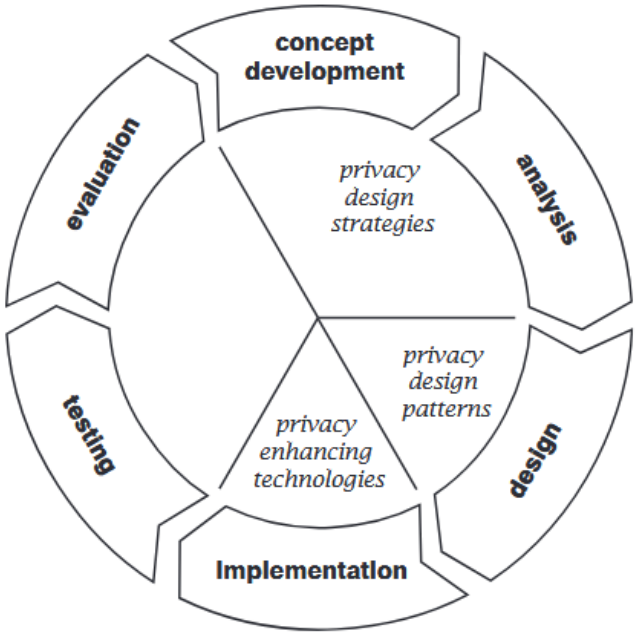
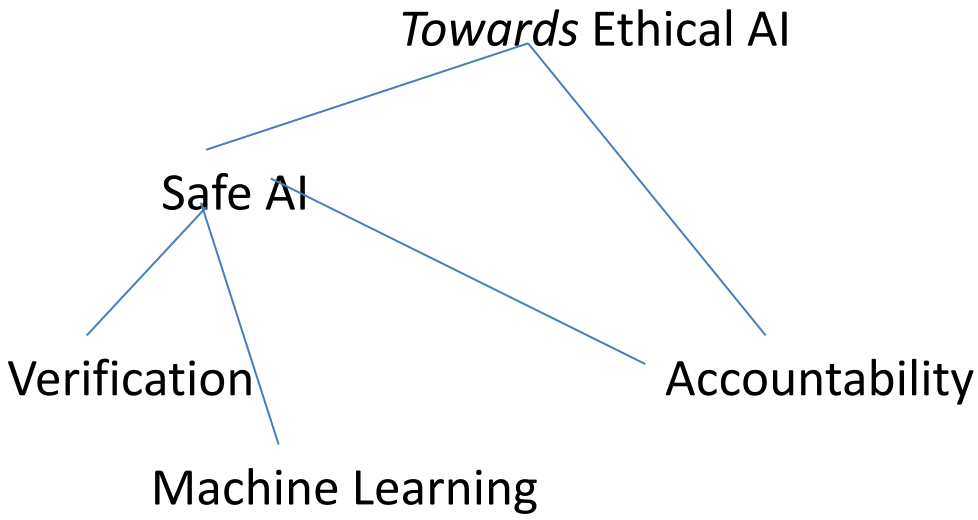


Fig. 1 The software development cycle and the relationship with strategies, patterns and technologies.



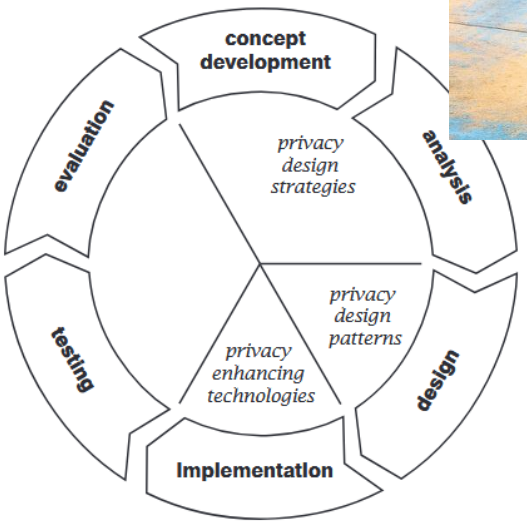
https://dtai.cs.kuleuven.be/projects/verilearn

In conclusion …











Images

pp. 1, 2, 60: https://storage.googleapis.com/english-cms-cwg-tw/article/201804/article-5ad9e57b828ca.jpg

Havasupai case images from Harmon, A. (2010). (see below)

https://upload.wikimedia.org/wikipedia/commons/c/c8/Jeremy_Bentham_by_Henry_William_Pickersgill_detail.jpg

https://upload.wikimedia.org/wikipedia/en/3/3d/John_Rawls.jpg By Harvard Gazette, Fair use, https://en.wikipedia.org/w/index.php?curid=24077224

https://media1.britannica.com/eb-media/36/97036-004-E9B37C21.jpg

http://www.greatthoughtstreasury.com/sites/default/files/Noddings%5B1%5D.gif

https://upload.wikimedia.org/wikipedia/commons/c/cc/Immanuel_Kant_%28portrait%29.jpg

http://1.bp.blogspot.com/_IoU3bEFUwWc/TLG4kYycezI/AAAAAAAAKds/1ZsDbuSCvhE/s400/Philippa+Foot.jpg

pp. 18, 53: from Awad et al. (2018).

http://cdn.cnn.com/cnnnext/dam/assets/130616121308-edward-snowden-getty-image-super-tease.jpg

http://cdn.nybooks.com/wp-content/uploads/2013/06/arendt_archive_1-071113.jpg

I heard the "unjust, but not unfair" example I mentioned along with slide 33 in Hellman, D. (2018). What is discrimination, when is it wrong and why? Keynote at FAT 2018. https://www.fatconference.org/2018/program.html

p. 38: http://www.amazon.com/Save-Everything-Click-Here-Technological/dp/1610393708

p. 39: De Raedt, L. & Meert, W. (2017ff.). ML Group Data Policy – Our Best Practice. KU Leuven, Belgium: Internal Document.

p. 40: https://pbs.twimg.com/media/B0FxelrCQAIozpJ.jpg, https://medium.com/the-establishment/even-doing-academic-research-on-video-games-puts-me-at-risk-1c48febb0c96

pp. 42ff.: https://ksassets.timeincuk.net/wp/uploads/sites/55/2017/07/fun-facts-about-the-incredibles-920x584.jpg, https://cdn-images-1.medium.com/max/2000/0*_yH7k0JyIcdEOqBt.jpg, https://www.rollingstone.com/wp-content/uploads/2018/06/oceans-eleven-movie-f4005da8-4e4b-4291-bcef-ff27294b2030.jpg, https://ksassets.timeincuk.net/wp/uploads/sites/55/2017/12/Oceans-8-poster-920x584.png

pp. 47, 61: https://static.adweek.com/adweek.com-prod/wp-content/uploads/2017/07/we-try-harder-hed-2017-640x360.jpg

p. 52: https://3c1703fe8d.site.internapcdn.net/newman/gfx/news/hires/2016/arerobotstak.jpg

p. 54: https://futureoflife.org/wp-content/uploads/2018/07/lethal-autonomous-weapons-pledge.jpg

The quote on p. 55 is from https://medium.com/fast-company/a-skeptics-guide-to-thinking-about-ai-4b5cfcadba09

pp. 59, 60: from Hoepman (2013/2014).

p. 60: https://www.bps.org.uk/sites/bps.org.uk/files/Policy/Policy%20-%20Images/Code%20of%20Ethics.jpg , https://humanrightshouse.org/wp-content/uploads/2018/01/17520.jpg

References

Aggarwal, A. (2018). The current hype cycle in Artificial Intelligence. KDNuggets. https://www.kdnuggets.com/2018/02/current-hype-cycle-artificial-intelligence.html

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment.

Nature, 563, 59-64.

Berendt, B. (2019). AI for the Common Good?! Pitfalls, challenges, and Ethics Pen-Testing. Paladyn. Journal of Behavioral Robotics, 10, 44-65. https://arxiv.org/abs/1810.12847

Berendt, B. & Preibusch, S. (2014). Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. Artificial Intelligence and Law, 22 (2), 175-209. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_preibusch_2014.pdf

Berendt, B. & Preibusch, S. (2017). Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop – and under the looking-glass. Big Data, 5(2). DOI: 10.1089/big.2016.0055.
https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_preibusch_2017_last_author_version.pdf

Bovens M. (2007). Analysing and assessing public accountability. A conceptual framework. European Law Journal, 13, 447–468.

Fisher, M. (2013). Full text of Snowden's new statement: I had 'the power to change people's fates'. The Washington Post, 12 July 2013.

https://www.washingtonpost.com/news/worldviews/wp/2013/07/12/full-text-of-snowdens-new-statement-i-had-the-power-to-change-peoples-fates/?utm_term=.599b71598f28

Harmon, A. (2010). Indian Tribe Wins Fight to Limit Research of Its DNA. The New York Times, 21 April 2010.
http://www.nytimes.com/2010/04/22/us/22dna.html

Hoepman J.-H. (2014). Privacy design strategies. Proc. ICT Systems Security and Privacy Protection - 29th IFIP TC 11 International Conference. pp. 446–459, 2014. Extended version available at https://arxiv.org/pdf/1210.6621.pdf

Morozov, E. (2013). To Save Everything, Click Here: The Folly of Technological Solutionism. New York, PublicAffairs.

Principles of International Law Recognized in the Charter of the Nürnberg Tribunal and in the Judgment of the Tribunal, 1950.
http://legal.un.org/ilc/texts/instruments/english/draft_articles/7_1_1950.pdf

Patel , N.A., & Elkin, G.D. (2015). Professionalism and Conflicting Interests: The American Psychological Association's Involvement in Torture. AMA Journal of Ethics, 17(10), 924-930. http://journalofethics.ama-assn.org/2015/10/nlit1-1510.html

Suomela, T., Chee, F., Berendt, B., Rockwell, G. (2018 / in press). Applying an Ethics of Care to Internet Research : Gamergate and Digital Humanities. To appear in Digital Studies / Le champ numérique. https://people.cs.kuleuven.be/~bettina.berendt/Papers/suomela_et_al_2018_last_author_version.pdf

Sweeney, L. (2013). Discrimination in Online Ad Delivery. Communications of the ACM, 56(5), 44-54. cited from earlier version available at SSRN: http://ssrn.com/abstract=2208240 or http://dx.doi.org/10.2139/ssrn.2208240

von Schirach, F. (2016). Terror. btb Verlag (eBook)Verlagsgruppe Random House. English version: Terror, Faber & Faber 2017.

Arendt: my own paraphrase of a key point made in Eichmann in Jerusalem: A Report on the Banality of Evil. Viking Press. 1963.

(The book began as an article for the New Yorker, which is available at https://www.newyorker.com/magazine/1963/02/16/eichmann-in-jerusalem-i)