Ethical Issues and Concerns in Digital Innovation

ISM-E2002

Kari Koskinen

Hadi Ghanbari

Department of Information and Service Management

Aalto University
School of Business

Image by Rosy from Pixabay

Session 5 – Imagining futures and AI ethics

Future studies

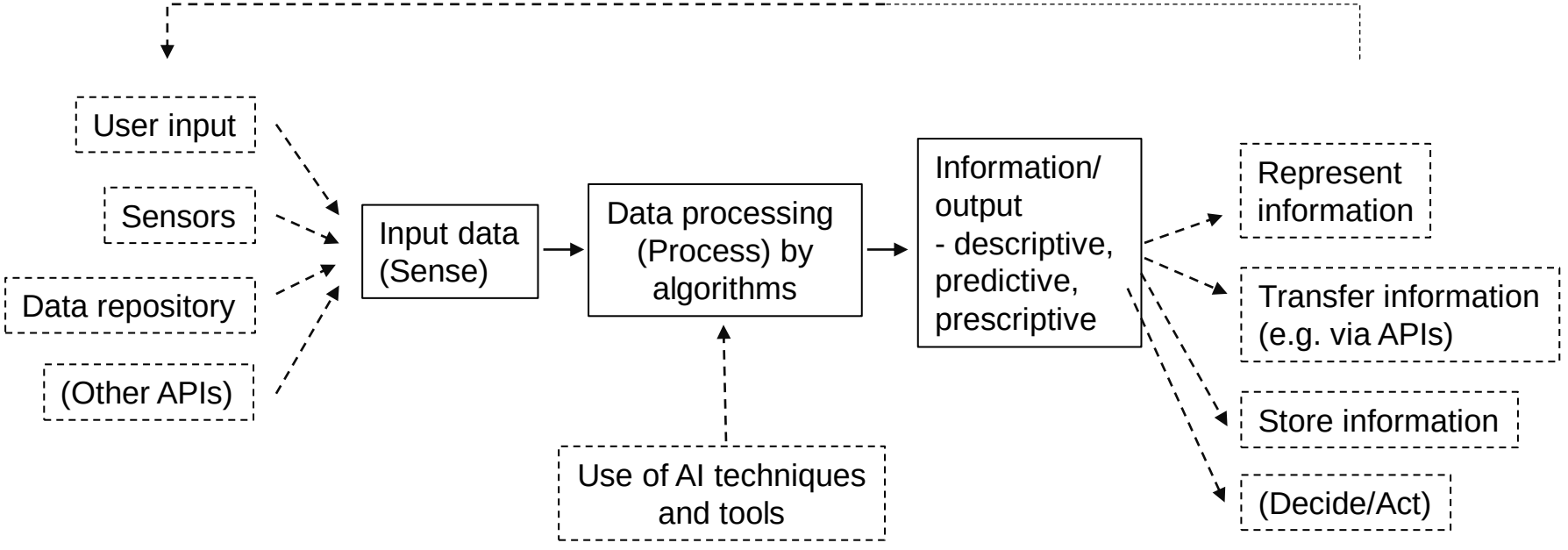
- Focus on technological futures – lack of consideration for the social and cultural, maintain existing ethical-political, social, and environmental dimensions
- Techniques:
  - Discovering the future
  - Future making
  - Futures imagined - Artefacts for the Future

Type of AftF	Embodied purpose	Questions arising
Type 1: AftF demonstrating future technology	This is what a technology will look like.	Where is it from? What would that world be like?
Type 2: AftF creating vicarious experiences	This is what it will be like.	How did we get here? What are the implications?
Type 3: AftF creating an intended impact	This is what the outcome might be.	How did we get here? How can we prevent it?
Type 4: AftF creating thought experiments	What if we pursued this?	Why is this happening? Do I want this? Is it right?

(Peter et al. 2020)

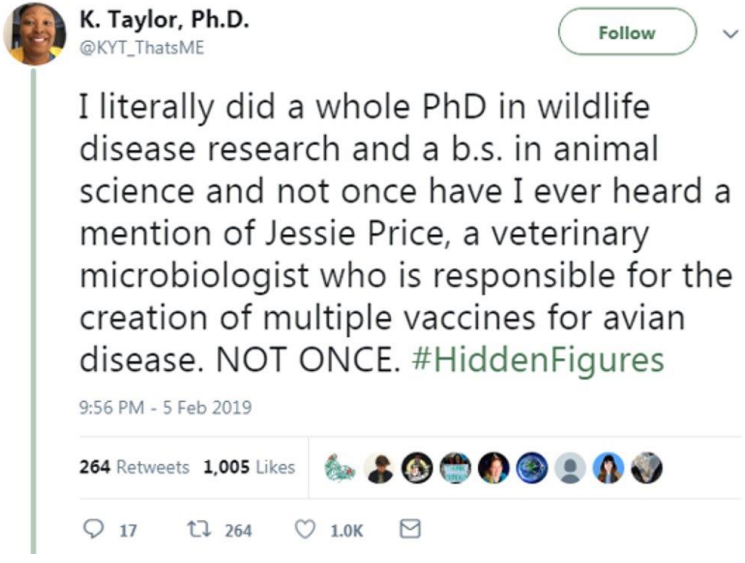
Data and algorithms

From data to information (and back)



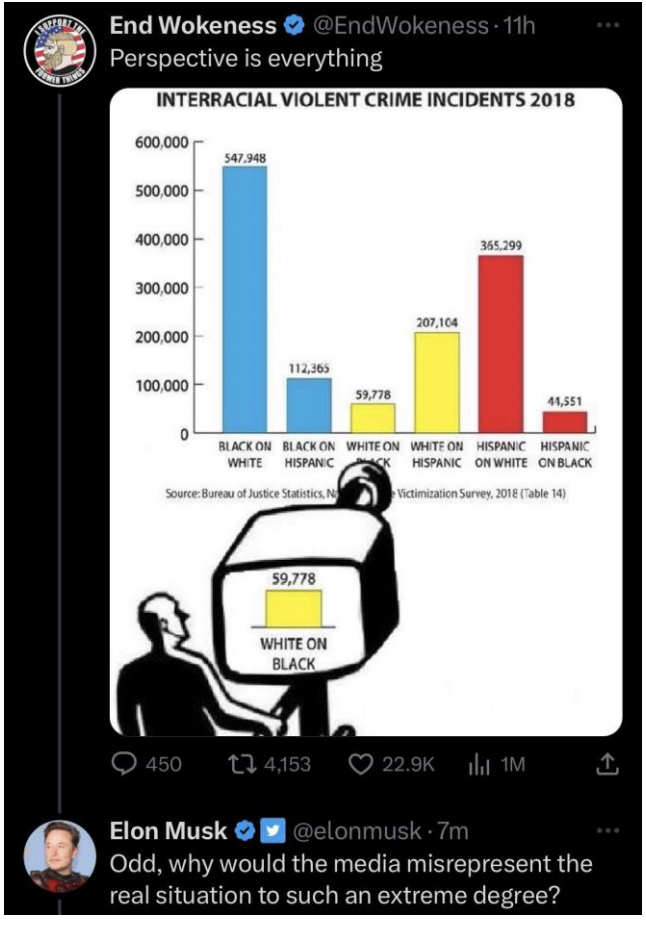
Semantics – meaning of data

- Semantics as the meaning or relationship of meanings of a sign or set of signs
- Occasionally difficult to infer meaning especially from unstructured data
- Meanings are given to data, often through negotiation) and represented in algorithms
- Who gets to decide and how does that fit with “reality”?
  - Raw data is an oxymoron (Bowker 2005)



Data presentations/visualisations

- How data is presented is subject to decisions
  - What is the message that is wanted to transfer?
    - What data are included?
    - What is the format?
    - What scales to use, which timeframe?
  - This does not mean that all presentations are false, simply that it is of importance to approach data and data presentations/ visualisations critically



Aalto University
School of Business

Example based on a thread of tweets by Kareem Carr, statistics PhD student: https://twitter.com/kareem_carr/status/1655614577445584897

Data presentations (cont.)

- ***Format:*** table vs. bar chart, count vs. percentage
  - Percentage of black people in the population 14.6%

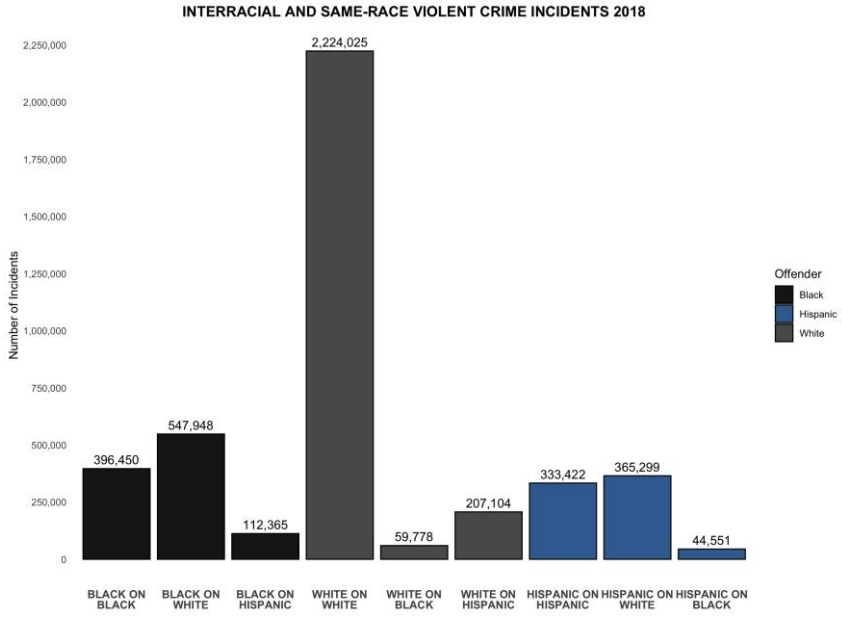
TABLE 14
Percent of violent incidents, by victim and offender race or ethnicity, 2018

Victim race/ ethnicity	Number of violent incidents	Offender race/ethnicity			
		Total	White[a]	Black[a]	Hispanic
White[a]	3,581,360	100%	62.1%*	15.3% †	10.2% †
Black[a]	563,940	100%	10.6 †	70.3*	7.9 †
Hispanic	734,410	100%	28.2 †	15.3 †	45.4*

- ***Organization of information:*** If most of the people are white and where victims are being selected purely at random, the vast majority of crimes committed by black offenders “should” involve white victims.

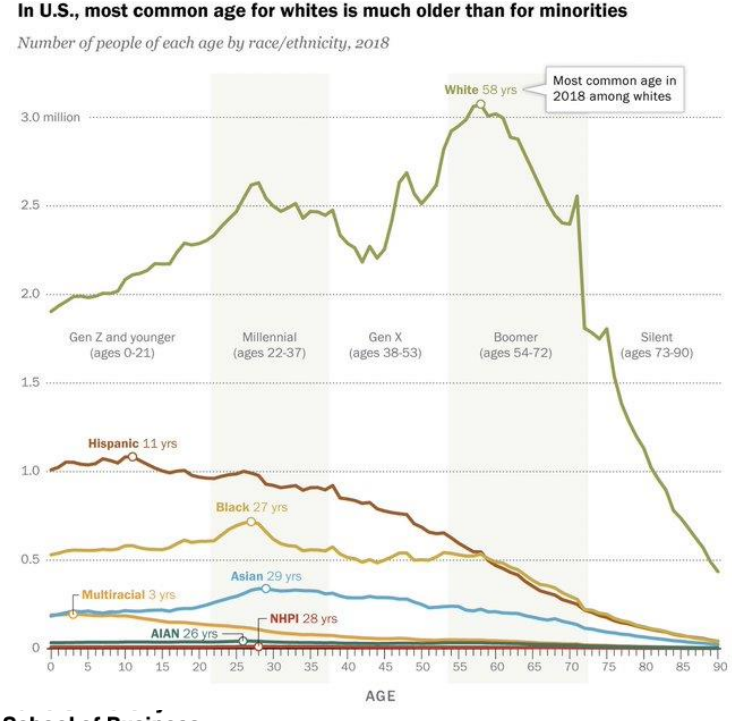
Data presentations (cont.)

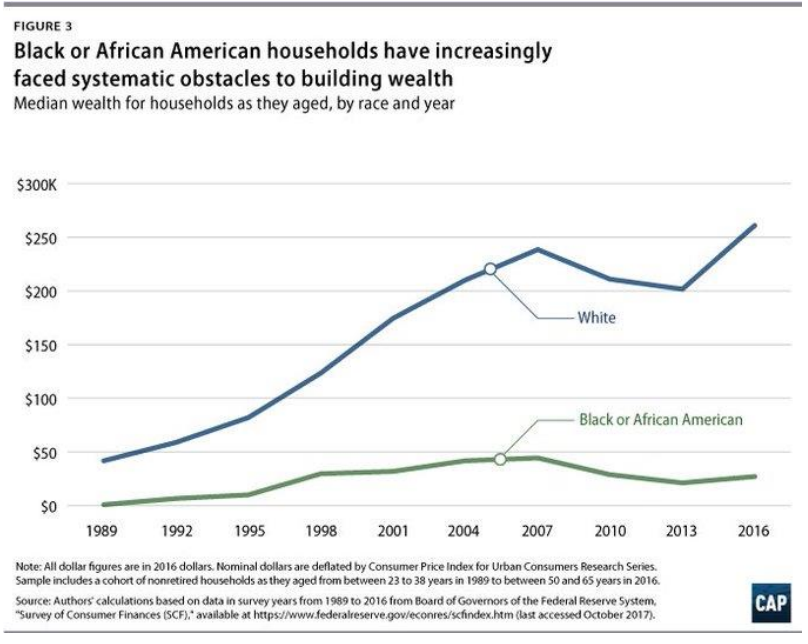
- ***Missing context***: by adding same-race crime, it becomes clear that the vast majority of violent crimes in America involve white victims and white offenders
  - If race is considered a big factor, reduction of crime numbers should start from there



Data presentations (cont.)

- ***Unfair comparison:*** age and wealth



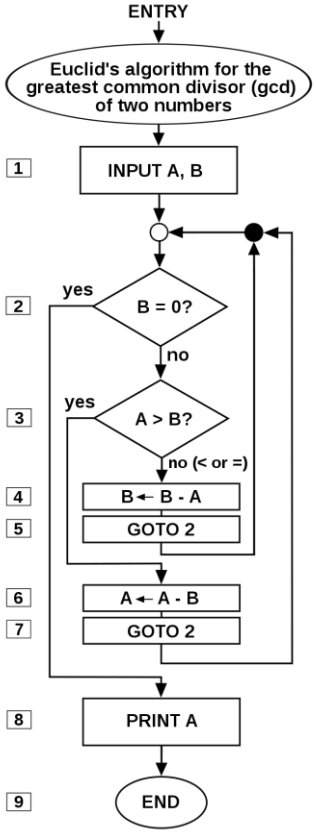


Data presentations (cont.)

- ***Misleading measure***: "Crime" is socially constructed by humans and when comparing social constructs between groups, there is a need to if the thing that is observed is socially constructed in exactly the same way for both groups
  - Do violent crime incidents have exactly the same probability of being reported and investigated regardless of race?
  - Do offenders have exactly the same probability of being arrested?
  - Do arrestees have exactly the same probability of being convicted?

Algorithms

- A procedure used for solving a problem or performing a computation
  - Algorithms act as a list of instructions that conduct specified actions step by step to process data
  - Data as a key ingredient for functioning algorithms
- Reflect values and objectives of their designers
  - Intended and unintended consequences
- Examples of algorithmic mistakes
  - Category mistakes: false positives and negatives
  - Process mistakes: mistakes in how a decision was made
- Calls for increased transparency and accountability
  - Tools and processes to identify, judge, and correct mistakes

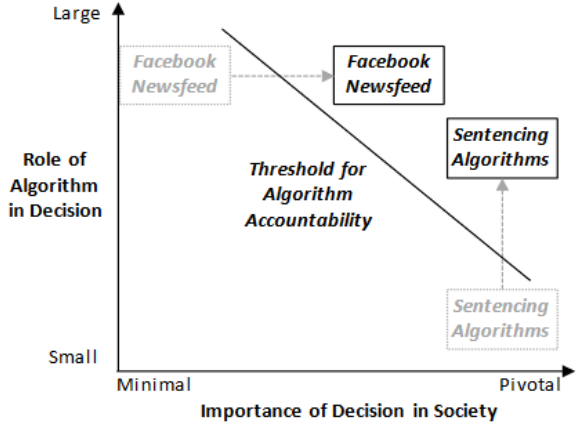


By Somepics - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=37859894

(Martin, 2019)

Humans and autonomy

- The use of algorithms and AI within them enables more automatization and machine autonomy
- Human role:
  - Human-in-the-loop: human has to intervene at some point of the process, for instance to confirm or change the outcome of an event or process
  - Human-on-the-loop: human monitors and can intervene the process
  - Human-out-of-the-loop: human out of the loop and letting the machines do all the learning and decision-making



(Martin, 2019)

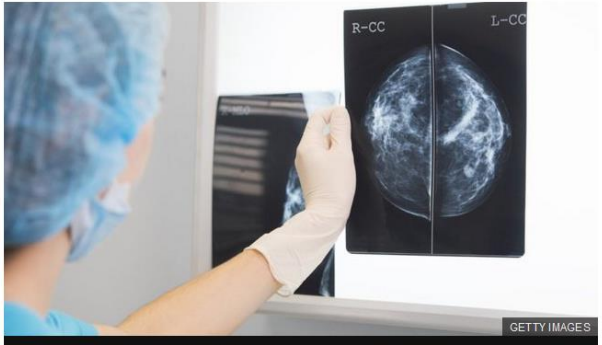
Artificial Intelligence (AI)

Examples of artificial intelligence (AI) use cases

AI 'outperforms' doctors diagnosing breast cancer

Fergus Walsh
Medical correspondent
@BBCFergusWalsh

2 January 2020 Share



AI was as accurate as two doctors working together

Artificial intelligence is more accurate than doctors in diagnosing breast cancer from mammograms, a study in the journal Nature suggests.

https://www.bbc.com/news/health-50857759

AlphaZero AI beats champion chess program after teaching itself in four hours

Google's artificial intelligence sibling DeepMind repurposes Go-playing AI to conquer chess and shogi without aid of human knowledge

▲ AlphaZero's victory is just the latest in a series of computer triumphs over human players since Computer programs have been able to beat the best IBM's Deep Blue defeated Garry Kasparov in 1997. Photograph: 18percentgrey / Alamy/Alamy

AlphaZero, the game-playing AI created by Google sibling DeepMind, has beaten the world's best chess-playing computer program, having taught itself how to play in under four hours.

https://www.theguardian.com/technology/2017/dec/07/alphazero-google-deepmind-ai-beats-champion-program-teaching-itself-to-play-four-hours

Photo: Margaret Minsky

Dartmouth workshop 1956:

The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

https://en.wikipedia.org/wiki/Dartmouth_workshop

Artificial intelligence

- Intelligence as the ability to learn, understand, to solve problems and to make decisions.
- AI combines computer science and robust datasets to enable problem-solving.
    - Utilised to make descriptions, prescriptions, and predictions based on input data
- AI as an umbrella term for various techniques such as machine learning and natural language processing
    - AI learn from data and modify their functioning

Machine Learning

- Teach a computer, through examples, to solve a problem
- ”I have a labelled set of photos of different plants. Build me a classifier that can tell me if an (unseen) photo contains a Tennessee purple coneflower or a Common Dandelion”

Symbolic AI

- Rule-based, hand-written logic. Similar to ”traditional” programming
- ”if credit rating is below this fixed threshold, approve the loan application. If not, reject it.”

Types of AI

- Narrow AI: intelligent systems that have been taught or learned how to carry out specific tasks without being explicitly programmed how to do so
- General AI: adaptable intellect found in humans, capable of learning how to carry out vastly different tasks or to reason about a wide variety of topics based on its accumulated experience (HAL in 2001 Space Odyssey, C3PO in Star Wars, Agent Smith in Matrix)
- Questions on self-awareness, having a “mind”, superintelligence..

https://twitter.com/kareem_carr/status/1640462020310077441

Common AI approaches for learning

- Supervised learning
  - Systems are fed data, which has been annotated to highlight the features of interest. Once trained, the system then applies these labels to new data.
- Unsupervised learning
  - There are no labels or correct outputs, and the algorithm isn't setup in advance to pick out specific types of data. The task is to discover the structure of the data: for example, grouping similar items to form “clusters”, or reducing the data to a small number of important “dimensions”.
- Reinforcement learning
  - Game-theoretic models that learn from a reward, basically going through a process of trial and error. Commonly used in situations where an AI agent like a self-driving car must operate in an environment and where feedback about good or bad choices is available with some delay.

AI design pitfalls and ethical challenges

AI design challenges - using cleaning robot (CR) as an example

- **Avoiding Negative Side Effects** - How to ensure that our CR will not disturb the environment in negative ways while pursuing its goals, e.g. by knocking over a vase because it can clean faster by doing so?
- **Avoiding Reward Hacking** - How to ensure that the CR won't game its reward function? Cover over messes with materials it can't see through.
- **Scalable Oversight** - How to ensure that the CR respects aspects of the objective that are too expensive to be frequently evaluated during training? E.g. do not throw out things that belong to someone.
- **Safe Exploration -** How to ensure that the CR doesn't make exploratory moves with very bad repercussions? For example, putting a wet mop in an electrical outlet
- **Robustness to Distributional Shift** - How to ensure that the cleaning robot recognizes, and behaves robustly, when in an environment different from its training environment?

(Amodei et al. 2016)

AI “bugs”

- **Space War:** Algorithms exploited flaws in the rules of the galactic videogame Elite Dangerous to invent powerful new weapons *(Scalable Oversight, Safe Exploration?)*
- **Body Hacking:** A four-legged virtual robot was challenged to walk smoothly by balancing a ball on its back. Instead, it trapped the ball in a leg joint, then lurched along as before (*Avoiding Reward Hacking*)
- **Goldilocks Electronics:** Software evolved circuits to interpret electrical signals, but the design only worked at the temperature of the lab where the study took place *(Robustness to Distributional Shift)*
- **Optical Illusion:** Humans teaching a gripper to grasp a ball accidentally trained it to exploit the camera angle so that it appeared successful—even when not touching the ball (*Avoiding Reward Hacking*)
- **Infanticide:** In a survival simulation, one AI species evolved to subsist on a diet of its own children *(Avoiding Negative Side Effects)*

https://www.wired.com/story/when-bots-teach-themselves-to-cheat/

AI societal challenges

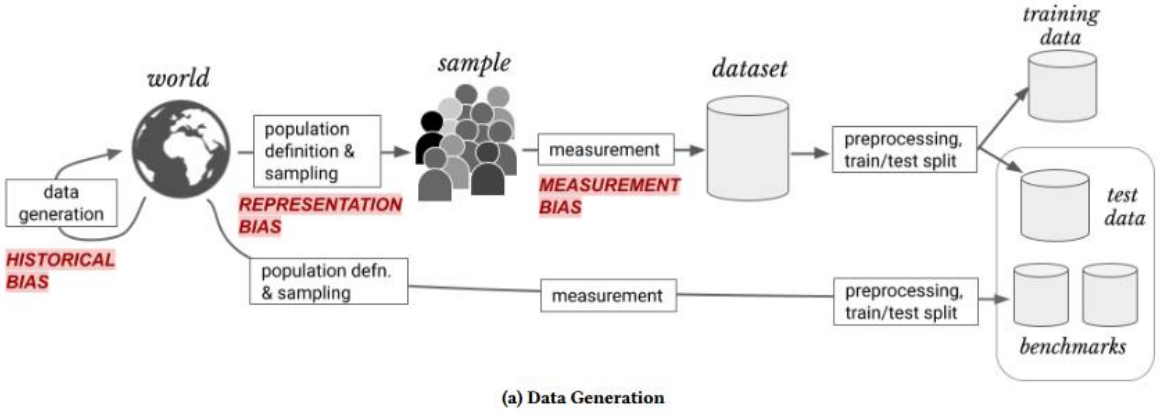
- **Privacy:** How can we ensure privacy when applying AI to sensitive data sources such as medical data?
- **Fairness:** How can we make sure AI systems don't discriminate?
- **Security:** What can a malicious adversary do to an AI system?
- **Abuse:** How do we prevent the misuse of AI systems to attack or harm people?
- **Transparency:** How can we understand what complicated ML systems are doing?
- **Policy:** How do we predict and respond to the economic and social consequences of AI?

(Amodei et al. 2016)

Data quality and bias

- Techniques such as ML are fundamentally based on statistical inference
  - Quality and amount of learning data set limits for ”intelligence”
- Biased data leads to biased intelligence
  - Garbage in, garbage out
    - System may receive new data and learn otherwise
  - Data to represent “what should be” and not “what is” (Hume’s guillotine)
    - Relevance for accuracy of decision-making?
- Various examples:
  - COMPAS
  - Amazon hiring algorithm
  - Disinfectant dispenser

Types of biases – data generation

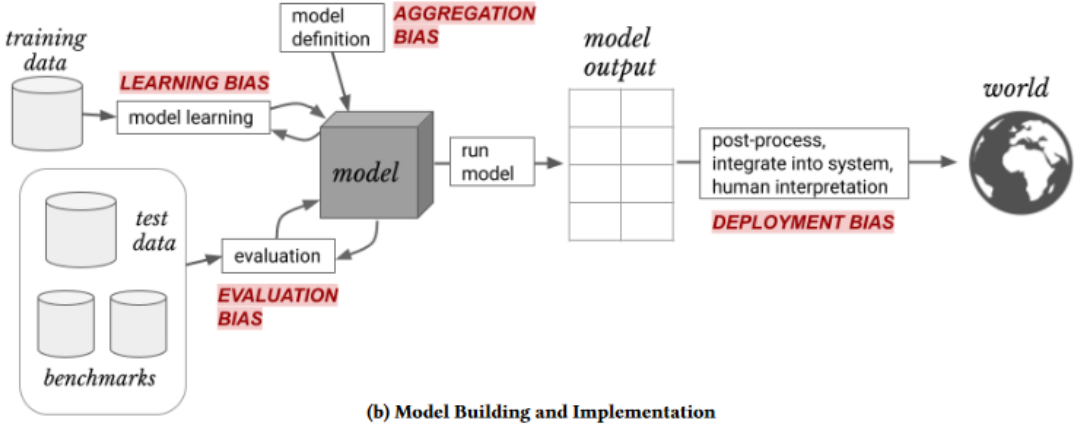


(a) Data Generation

- Historical bias: represents the world as it is or was leads to a model that produces harmful outcomes, such as reinforcing a stereotype of a group
- Representation bias: sample underrepresents some part of the population and fails to generalize well for a subset of the use population.
- Measurement bias: errors in choosing, collecting, or computing features and labels to use in a prediction problem (e.g. a proxy poorly reflects a construct or generated differently across groups).

From: Suresh & Guttag 2021
(https://arxiv.org/pdf/1901.10002.pdf)

Types of biases – implementation



(b) Model Building and Implementation

- Aggregation bias: a one-size-fits-all model is used for data containing underlying groups that should be considered differently
- Learning bias: choices regarding the model amplify performance disparities (e.g. selection of objective function harms another)
- Evaluation bias: benchmark data for a task not representing the use population
- Deployment bias: mismatch between the problem a model is intended to solve and for what it is actually used

From: Suresh & Guttag 2021
(https://arxiv.org/pdf/1901.10002.pdf)

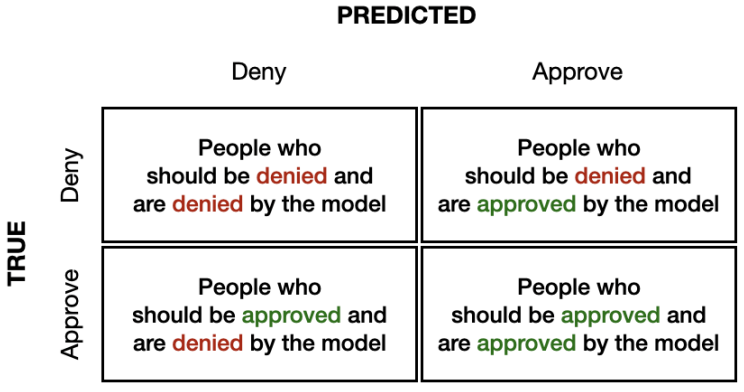
Model (un)fairness

- Four criteria
  - ***Demographical parity:*** model is fair if the composition of people who are selected by the model matches the group membership percentages
  - ***Equal opportunity:*** the proportion of people who should be and are correctly selected by the model ("positives") is the same for each group (true positive rate)
  - ***Equal accuracy:*** the percentage of correct classifications should be the same for each group
  - ***Group unaware fairness:*** removal of all group membership information (also proxy data) from the dataset (e.g. age data to aim to make the model fair to different age groups)

https://www.kaggle.com/code/alexisbcook/ai-fairness

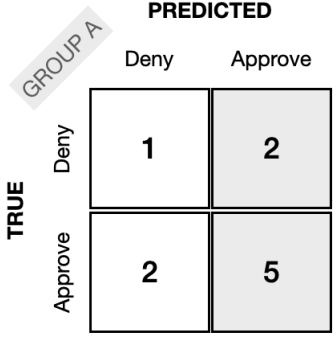
Evaluating fairness

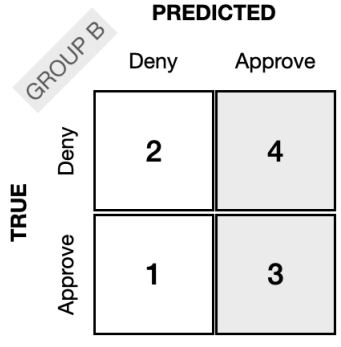
- Confusion matrix – tool to assess a model's performance across groups (apart from group unaware)
- In practice, it is not possible to optimize a model for more than one type of fairness
  - Impossibility theorem of machine fairness (Sa[illegible])



Demographic parity

20 applicants (50% from **Group A**)
14 approvals (50% from **Group A**)





https://www.kaggle.com/code/alexisbcook/ai-fairness

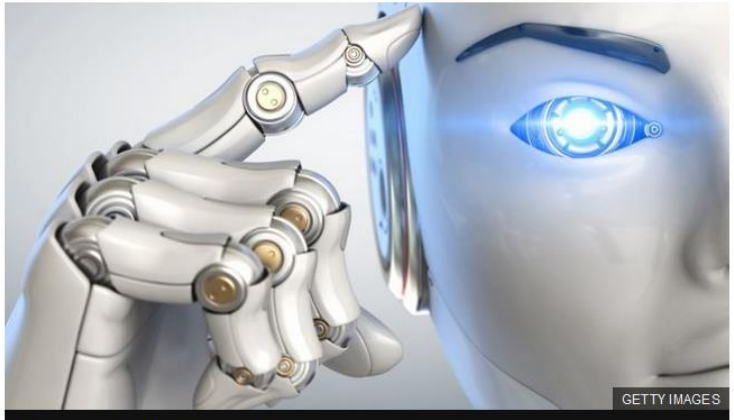
Importance of theory

- Black-boxing technology
  - Complex, proprietary algorithms often trade secrets
  - Occasionally difficult, if not impossible, to know exactly why a complex AI system reached the result it did, even if the system as such open
- “If it works, does it matter?”
  - Explainability of AI
  - Understanding how certain decisions are made

Can we trust AI if we don't know how it works?

By Marianne Lehnis
Technology of Business reporter

15 June 2018 Share

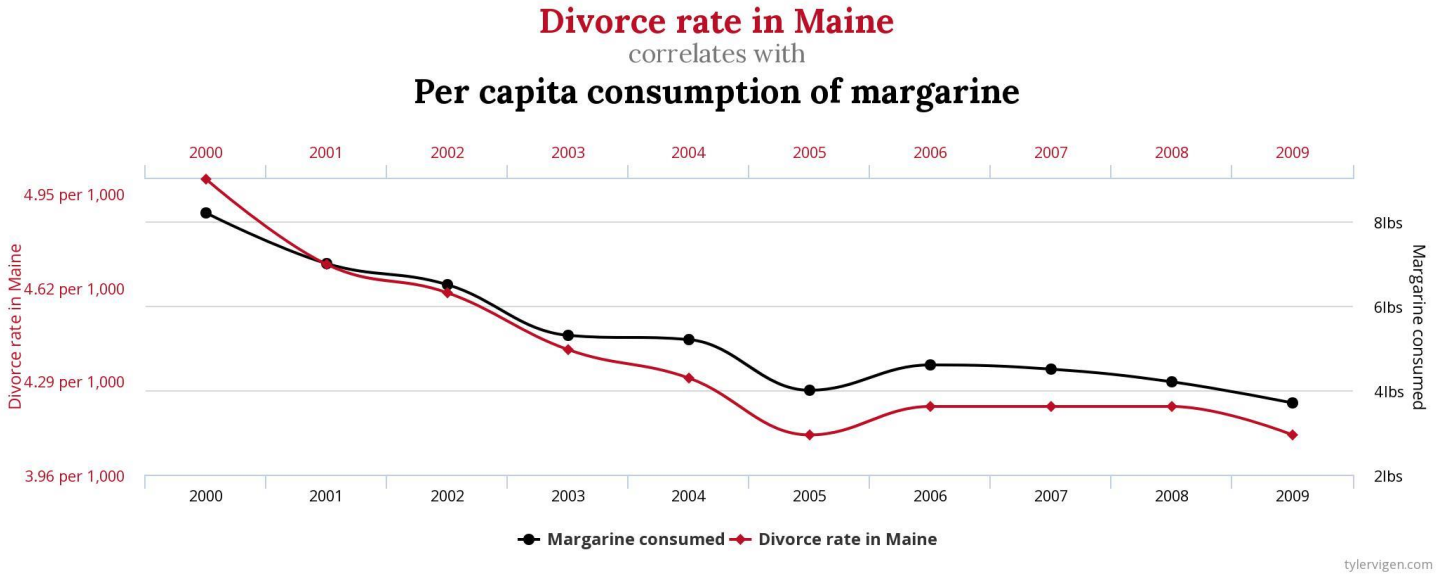


What's it thinking? Will AI become too clever for us?

We're at an unprecedented point in human history where artificially intelligent machines could soon be making decisions that affect many aspects of our lives. But what if we don't know how they reached their decisions? Would it matter?

https://www.bbc.com/news/business-44466213

Importance of theory – finding patterns



https://www.tylervigen.com/spurious-correlations

Approaches to Ethical AI

AI - design principles

- Various institutions and companies have created their own set of AI (or related field) ethics principles
    - Algorithm Watch's inventory contains 167 guidelines
- Different guidelines have same/similar principles
    - Beneficience, non-maleficiency, autonomy, justice, and explicability (Floridi & Cowls, 2019)
- Munn (2022): principles/practice gap
    - Principles meaningless (contested), isolated (larger than technical issues), toothless (recommendations, voluntary commitment)
    - Difficult to apply and have no impact to developers

DataforGood
Serment d'Hippocrate pour Data Scientist (Hippocratic Oath for Data Scientists)
France | civil society
Voluntary commitment

Telia Company
Telia Company Guiding Principles on trusted AI ethics
Sweden | private sector
Voluntary commitment

Deep Mind
Saftey and Ethics
United States | private sector
Voluntary commitment

Machine Intelligence Research Institute
PDF: The Ethics of Artifical Intelligence
United States | academia
Recommendation

Icelandic Institute for Intelligent Machines
IIIM's Ethics Policy
Iceland | civil society
Voluntary commitment

OP Financial Group
OP Financial Group's ethical guidelines for artificial intelligence
Finland | private sector
Voluntary commitment

https://inventory.algorithmwatch.org

EU: ethics guidelines for trustworthy AI

- **Human agency and oversight:** AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured
- **Technical Robustness and safety:** AI systems need to be resilient and secure. They need to be safe, ensuring a fall-back plan in case something goes wrong, as well as being accurate, reliable and reproducible.
- **Privacy and data governance:** besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.
- **Transparency:** the data, system and AI business models should be transparent. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system and must be informed of the system's capabilities and limitations.

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

EU: ethics guidelines for trustworthy AI

- **Diversity, non-discrimination and fairness:** Unfair bias must be avoided as it could have multiple negative implications. AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.
- **Societal and environmental well-being:** AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Social and societal impact should be carefully considered.
- **Accountability:** Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Moreover, adequate and accessible redress should be ensured.

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

EU’s ethical principles for AI research

1. **Respect for Human Agency:** human beings must be respected to make their own decisions and carry out their own actions, maintain their autonomy, dignity and freedom

2. **Privacy and Data governance:** people have the right to privacy and data protection

3. **Fairness:** people given equal rights and opportunities and should not be dis-/advantaged undeservedly

4. **Individual, Social and Environmental Well-being:** AI systems should contribute to, and not harm, individual, social and environmental wellbeing

5. **Transparency:** the purpose, inputs and operations of AI programs should be knowable and understandable to its stakeholders

6. **Accountability and Oversight:** humans should be able to understand, supervise and control the design and operation of AI based systems. Actors involved in their development or operations are responsible for the functioning of these applications and their consequences.

https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

Implementation of ethics in the development and use of AI systems

- Use of checklists
- Advice on how to implement the guidelines in system development, e.g. actions to take at each step of the development process

- Similarly, advice on how to implement the guidelines for use of the system, ranging from project management and acquisition of the system to its deployment, implementation and monitoring
- Finnish Center for Artificial Intelligence (FCAI) ethics exercise tool for research

Specification of Objectives against Ethical Requirements	Yes	No (how potential risks will be mitigated?)
Respect for Human Agency		
End-users and others affected by the AI system are not deprived of abilities to make all decisions about their own lives, have basic freedoms taken away from them,		
End-users and others affected by the AI system are not subordinated, coerced, deceived, manipulated, objectified or dehumanized, nor is attachment or addiction to the system and its operations being stimulated.		
The system does not autonomously make decisions about vital issues that are normally decided by humans by means of free personal choices or collective deliberations or similarly significantly affects individuals,		
The system is designed in a way that give system operators and, as much as possible, end-users the ability to control, direct and intervene in basic operations of the system (when relevant)		
Privacy & Data Governance		
The system processes data in line with the requirements for lawfulness, fairness and transparency set in the national and EU data protection legal framework and the reasonable expectations of the data subjects.		
Technical and organisational measures are in place to safeguard the rights of data subjects (through measures such as anonymization, pseudonymisation, encryption, and aggregation).		

https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

Human-Centred Design (HCD) for AI

- HCD captures the human perspective in different steps of a design process
  - Draws from participatory action research
  - Seek to understand the design problem/issue from the stakeholders/users/community perspective while also enabling participation in the design stage and after by collecting feedback etc.
- One approach (based on Shankar & Cook 2021, Google):
  - Define the problem by understanding people's needs and wants
  - Evaluate if AI is required in the proposed solution
  - What kinds of harms may the use of AI lead to
  - Use prototypes (non-AI also) to obtain feedback from people
  - Provide channels to challenge the outcomes produced by the system
  - Enable human monitoring, build safety measures and options to flag issues

https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

Model Cards

- Short documentation of key information regarding an AI model (Mitchell et al. 2019)
  - Objective to increase transparency on how the model functions (e.g. GPT-3)
- Should contain information such as
  - Model details, information about the model (person/organization developing model, model date, version, fairness constraints, where to send questions or comments)
  - Intended users and use cases, also out-of-scope uses cases
  - Factors such as demographic or phenotypic groups, environmental conditions, technical attributes
  - Metrics, should be chosen to reflect potential real-world impacts of the model (model performance measures, decision thresholds)
  - Evaluation data, details on the dataset(s) used for the quantitative analyses.
  - Training data, may not be possible to provide in practice
  - Ethical considerations

Responsible AI: Google

Case questions

1. Why is it difficult to implement ethical principles?
2. How well is Google doing in terms of its ethical approach towards AI? What is good, and what kinds of improvements could be made?
3. In your view, is Google serious or engaging in (AI) ethics-washing?

Case questions

1. What kinds of data might the AI-based lending decision solution require?
2. How might the individual data items contain or lead to bias?
3. It is stated that “ethical challenges are not just design flaws” and “technical solutions are not enough to fix the ethical issues”. What is meant with this and what does this mean in relation to the lending decision solution?