Image by DALLE2

THE UNIVERSITY OF MELBOURNE

Week 6/S1/2023

Transparency: Decisions & Processes

Marc Cheong

School of Computing and Information Systems
& Centre for AI & Digital Ethics
The University of Melbourne
marc.cheong [at] unimelb.edu.au

Learning Outcomes

1. Distinguish between transparency and explainability, closely-related concepts in AI ethics.
2. Understand how automated decision-making (ADM) systems require transparency at every stage.
3. Understand how tech companies approach the issue of transparency, as well as concerns that are raised by stakeholders, in areas where AI systems are deployed: e.g., social media ads, criminal justice, generative AIs.
4. Understand how big data research can - either positively or negatively - affect their data subjects, and why transparency is important when conducting such studies.

Related Reading

This module has two readings corresponding to the two broad themes within (plus an optional study).

Recap: screenshot below ☺

1. Experiments in Social Media
   Toby Walsh. *AI Magazine, 40(4), 74-77. 2019. Archived copy by the author in arXiv [cs.CY].*
   This is an interesting piece which highlights the dangers how studies/experiments on social media can contribute to "challenges[,] as even small effects when multiplied by a large population can have a significant impact". Experimental "interventions increased turnout by about 340,000 additional votes ... around 0.5% of the total number of votes cast" (Walsh, 2019) in a Facebook experiment to encourage voting in the 2010 US Elections. The author highlights the issues of (non)transparency in AI research especially on such a large scale.

2. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead
   Cynthia Rudin. *Nature Machine Intelligence*, 1, 206–215. 2019. *Archived copy by the author in arXiv [stat.ML].*
   The second reading for this week focuses on the pervasiveness of "black box machine learning models" (Rudin, 2019) and how their implementation -- in, say, decision making for criminal justice -- can be mired by a "lack of transparency and accountability of predictive models ... [with] severe consequences" (Rudin, 2019). We need to look at the processes and the decision making aspects behind the development and deployments of these systems. From a computer scientist's lens, we should ask questions like 'should we even use it on people?' rather than 'can we optimise it?'

Additional readings.

If you find this module interesting, you might want to check out the brief history of the Cambridge Analytica controversy on Wikipedia here.

Outline

1. Transparency: what's it all about.
2. Automated decision-making (ADM) systems: transparency at every stage?
3. Current issues in transparency and stakeholders' concerns,
   I. social media advertising systems
   II. criminal justice AI systems
   III. Generative AIs!
4. Data science research and transparency: the cases of Cambridge Analytica and Covid19 Trends.

THE UNIVERSITY OF MELBOURNE

In the spirit of the 'flipped classroom' and making lectures engaging (as with Simon), there will be sweets distributed in class as well.

Simon's 'cue' is to use Freddo frogs. Our cue is to spot the 'like' icon 👍

(also for engaging discussions, you will automatically get sweets, if you choose).

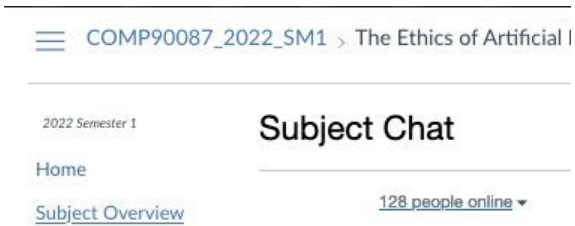
Transparency: What's it all about?

Image by DALLE2

Audience activity I [~5 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.

Incentive:

Image source: Cadbury

What is transparency?

What does it mean to you?

What does it mean when you say 'this AI system does things transparently'?

Image source: Eliiza

A note on nomenclature (and focus)

Some academics use the term transparency to describe the “inner workings” of algorithms (e.g. Loi et al, 2020: https://link.springer.com/article/10.1007/s10676-020-09564-w)

To clarify, the focus of this module is on the *overarching processes and decision-making* of an algorithmic *system.*

We include the technical considerations about interpretability of the algorithms themselves, counterfactual analysis, etc under the banner of *explainability*.

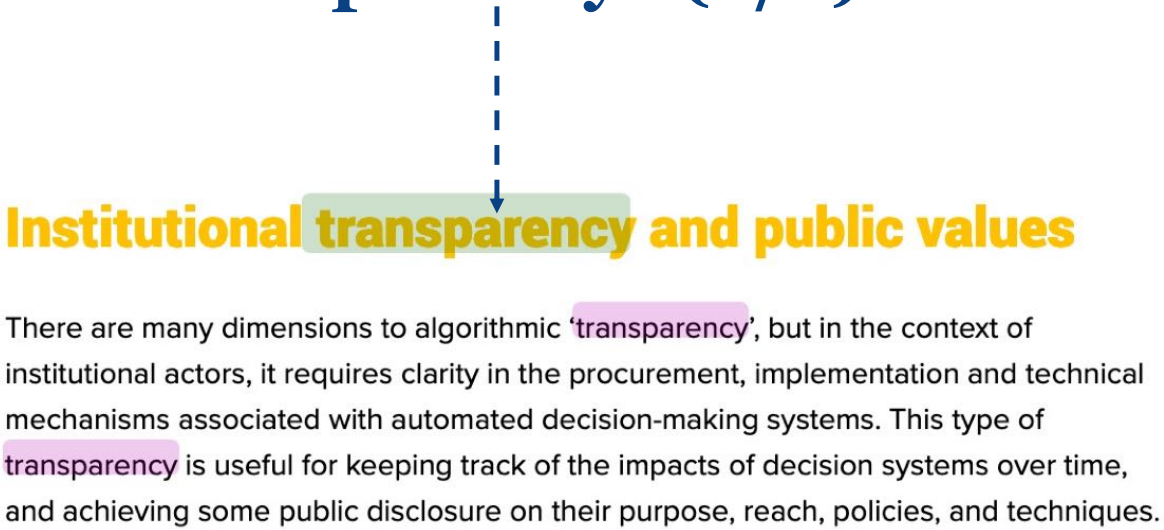
(Our very own Tim Miller is an expert in this field).

Transparency? (1/2)

As more decisions become automated and processed by algorithms, these processes become more opaque and less accountable, with risks of secret profiling and illegal discrimination. For Rotenberg, “at the core of modern privacy law is a single goal: to make transparent, the automated decisions that impact our lives.” He sees “algorithmic transparency”, the principle that data processes which impact individuals be made public, as the next stage in the development of transparency law, internet law and privacy law. The lack of algorithmic transparency in the current internet ecosystem poses a crucial challenge to defending fundamental human rights online, ranging from privacy and freedom of expression to security. In addition to algorithmic transparency, Rotenberg pointed to other emerging issues which need to be examined, notably the increasing access to drones and robots and the need for their registration.

UNESCO news release, quoting Marc Rotenberg: https://en.unesco.org/news/privacy-expert-argues-algorithmic-transparency-crucial-online-freedoms-unesco-knowledge-cafe

Transparency? (2/2)



Institutional transparency and public values

There are many dimensions to algorithmic 'transparency', but in the context of institutional actors, it requires clarity in the procurement, implementation and technical mechanisms associated with automated decision-making systems. This type of transparency is useful for keeping track of the impacts of decision systems over time, and achieving some public disclosure on their purpose, reach, policies, and techniques.

Jake Goldenfein, 'Algorithmic Transparency and Decision-Making Accountability: Thoughts for buying machine learning algorithms' in Office of the Victorian Information Commissioner (ed), Closer to the Machine: Technical, Social, and Legal aspects of AI (2019). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3445873

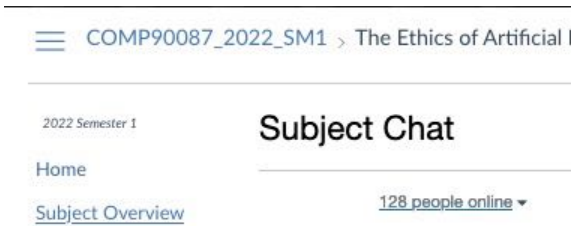
Example: Hypothetical Thought Experiment. Deep Learning and Your Grades!

"What if, for this unit, we decide your final grade based on a **state-of-the art deep learning-based estimator/predictor** that will come up with a final grade based on **1,000 factors** - ranging from how many Youtube videos you watch about 👍 ethics, to your activity in weekly discussion activities, to your typing speed when asked to comment on discussion boards, to your on-Zoom reactions to Simon and the tutors when they taught you the basics of ethics, to how much you laugh at Marc's internet memes in the modules, etc...
The final mark predictor is so advanced, **it has been audited by NASA, Google, and by 10 Nobel Prize winners**! However, since some of the **tensor-based algorithms used are proprietary to NVidia**, who sponsored the array of Geforce RTX4090s used for deep learning, they can't be revealed in public.
Also, the decision made by the predictor is **final**."

Audience activity II [5-10 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.

Incentive:

Image source: Cadbury

What is wrong with this thought experiment?

"What if, for this unit, we decide your final grade based on a **state-of-the art deep learning-based estimator/predictor** that will come up with a final grade based on **1,000 factors** - ranging from how many Youtube videos you watch about ethics, to your activity in weekly discussion activities, to your typing speed when asked to comment on discussion boards, to your on-Zoom reactions to Simon and the tutors when they taught you the basics of ethics, to how much you laugh at Marc's internet memes in the modules, etc...
The final mark predictor is so advanced, **it has been audited by NASA, Google, and by 10 Nobel Prize winners**! However, since some of the **tensor-based algorithms used are proprietary to NVidia**, who sponsored the array of Geforce RTX4090s used for deep learning, they can't be revealed in public.
Also, the decision made by the predictor is **final**."

Image source: HowToGeek / Imagentle / Shutterstock

Reflection/Some discussion points.

Process: who governs the selection of ML models/training data?
What vendors are given preference – e.g. why Google (not AWS)?
e.g why Tensorflow and Nvidia?
Any conflicts of interest? Any feedback loops?
Why is the whole system shrouded in secrecy?
Who audited it? Can I see the source code/design schematics/rationales?
Did I even sign up for this?

Decisions: can we challenge them? Can I take this to court?
Who sanctioned this to be official?
Is this another *Cambridge Analytica*?

Image source: The Pokemon Company

Audience activity II ½ [5-10 mins]

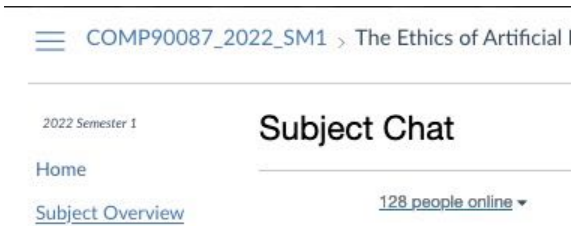
Let's help you practice philosophical argument!

SIMON USES PHILOSOPHICAL ARGUMENT.

Audience activity II ½ [5-10 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.
Incentive:

Image source: Cadbury

"What if, for this unit, we decide your final grade based on a **state-of-the art deep learning-based estimator/predictor** that will come up with a final grade based on **1,000 factors** - ranging from how many Youtube videos you watch about ethics, to your activity in weekly discussion activities, to your typing speed when asked to comment on discussion boards, to your on-Zoom reactions to Simon and the tutors when they taught you the basics of ethics, to how much you laugh at Marc's internet memes in the modules, etc...
The final mark predictor is so advanced, **it has been audited by NASA, Google, and by 10 Nobel Prize winners**! However, since some of the **tensor-based algorithms used are proprietary to NVidia**, who sponsored the array of Geforce RTX4090s used for deep learning, they can't be revealed in public.
Also, the decision made by the predictor is **final**."

Argue FOR (in support 👍) the deployment of this, based on utilitarianism.

Argue AGAINST (no way! ❌) the deployment of this, based on Kant's Duty Ethics

(1st CI: 'will' that everyone does this / 2nd CI: ends not means).

Automated decision-making (ADM) systems: *Transparency at every stage?*

Image by DALLE2

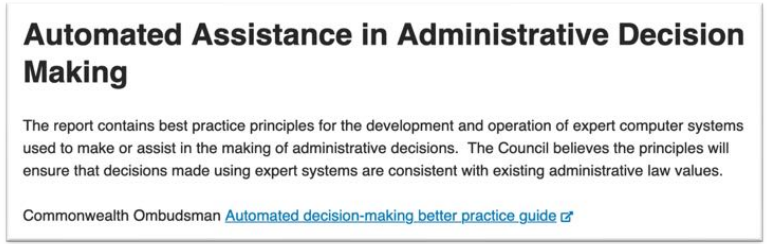
Image source: CAPCOM / Cinemablend.

Disclaimer: I am not a lawyer

The information provided in this mini-lecture is summarized from various sources to explain how transparency is a requirement not only for the algorithms, but also the contexts surrounding their implementation.

This lecture won't make you an expert in administrative decision-making ☺

AADM



Administrative Review Council, *Automated Assistance in Administrative Decision Making: Report to the Attorney General* (Report No 46, November 2004) (‘*2004 Report*’)

https://www.ag.gov.au/legal-system/publications/report-46-automated-assistance-administrative-decision-making-2004

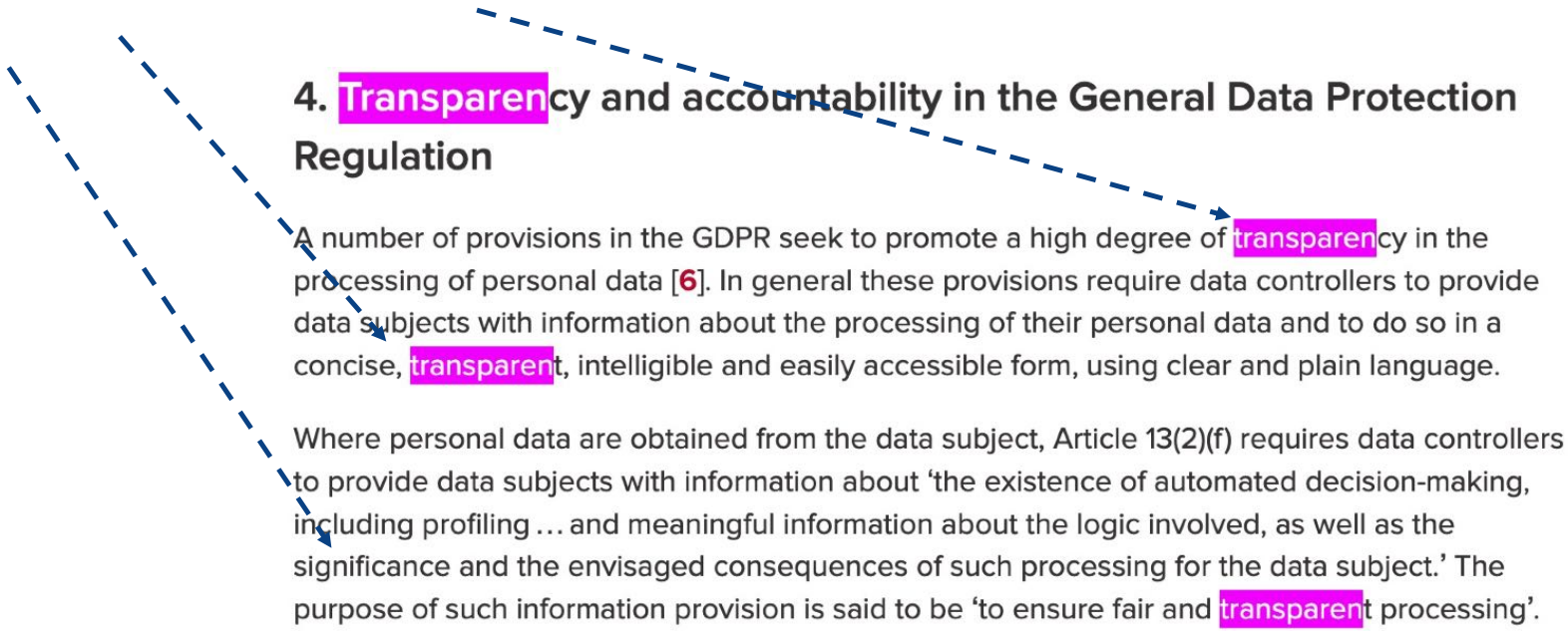
The 2004 Report was ahead of its time (emphases below are mine)

P27: “**Safeguards built into the system are only asking relevant questions, telling customers why questions are being asked (which makes the decision-making process more transparent)** and recording and explaining to a customer the reason for a decision”

P43: “Expert systems’ ability to provide an **audit trail of the administrative decision-making processes they are involved in** is important to the administrative law values of transparency, fairness and efficiency. “

P45: “A good system of internal review is one which **is transparent in process and affords a quick, inexpensive and independent review of decisions**. Such a system is beneficial both to applicants and agencies”. (citing Administrative Review Council 2000)

GDPR?



4. Transparency and accountability in the General Data Protection Regulation

A number of provisions in the GDPR seek to promote a high degree of transparency in the processing of personal data [6]. In general these provisions require data controllers to provide data subjects with information about the processing of their personal data and to do so in a concise, transparent, intelligible and easily accessible form, using clear and plain language.

Where personal data are obtained from the data subject, Article 13(2)(f) requires data controllers to provide data subjects with information about 'the existence of automated decision-making, including profiling … and meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.' The purpose of such information provision is said to be 'to ensure fair and transparent processing'.

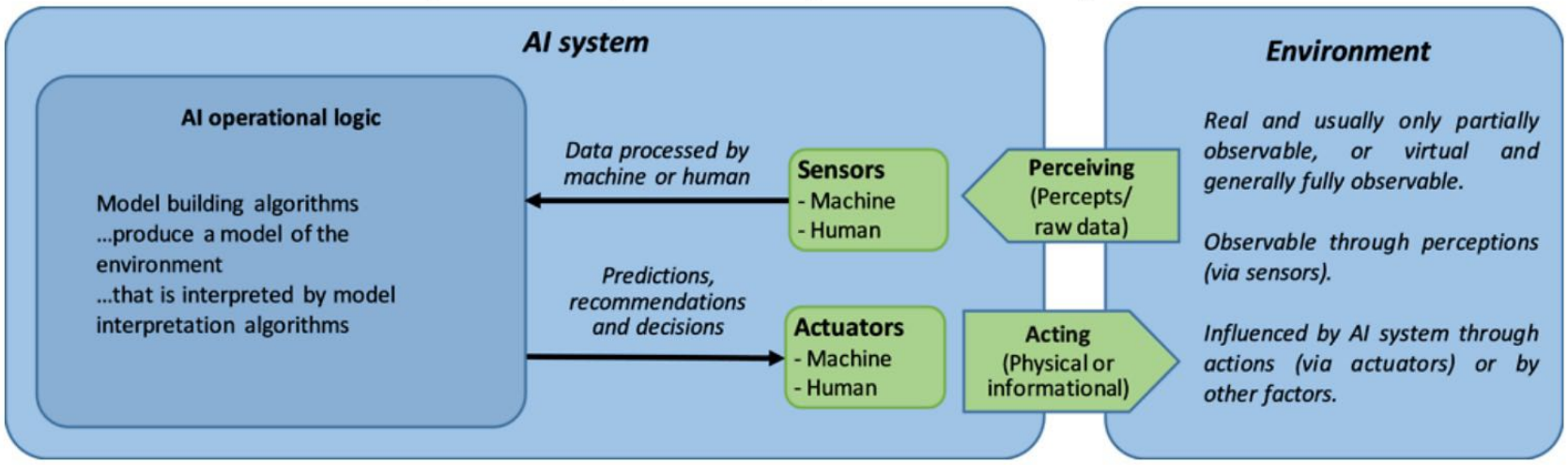
Christina Blacklaws, 'Algorithms: transparency and accountability', Phil. Trans. R. Soc. A.3762017035120170351. (2018). https://royalsocietypublishing.org/doi/10.1098/rsta.2017.0351

High-level view of AI systems (OECD, 2019)

Source: OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://doi.org/10.1787/eedfee77-en.

Figure: https://www.oecd-ilibrary.org/sites/8b303b6f-en/index.html?itemId=/content/component/8b303b6f-en#figure-d1e976

Figure 1.3. **A high-level conceptual view of an AI system**



Source: As defined and approved by AIGO in February 2019.

Elements for ML systems (Google Inc, from Scully et al 2015)

Source: https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning - adapted from Scully et al (2015) https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

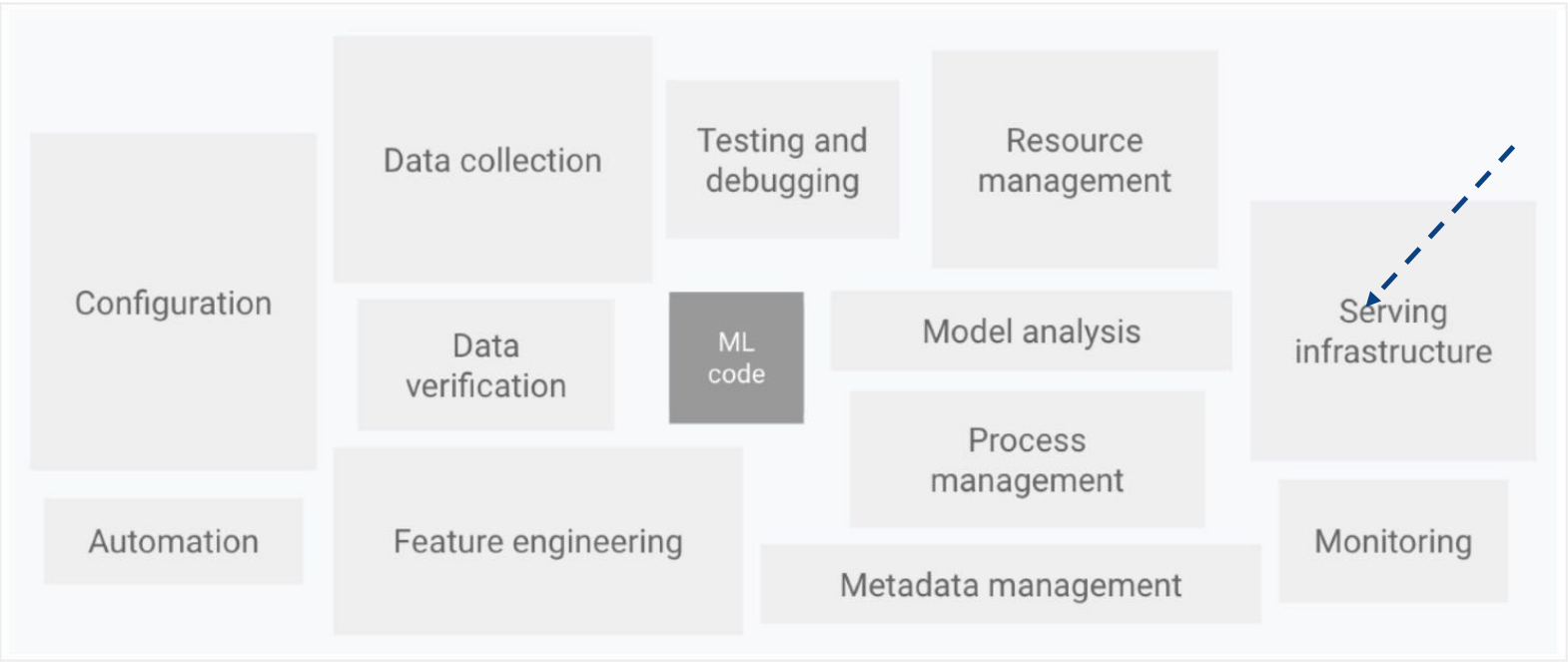
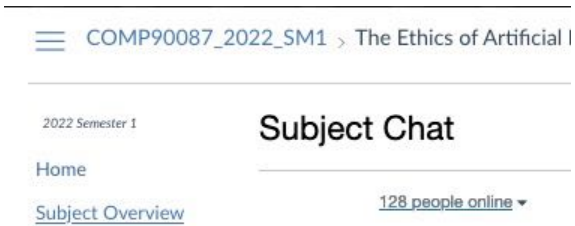


Figure 1. Elements for ML systems. Adapted from Hidden Technical Debt in Machine Learning Systems.

Audience activity III [5-10 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.

Incentive:

Image source: Cadbury

All other parts of a typical ML system can be made more transparent. Can you comment on how?

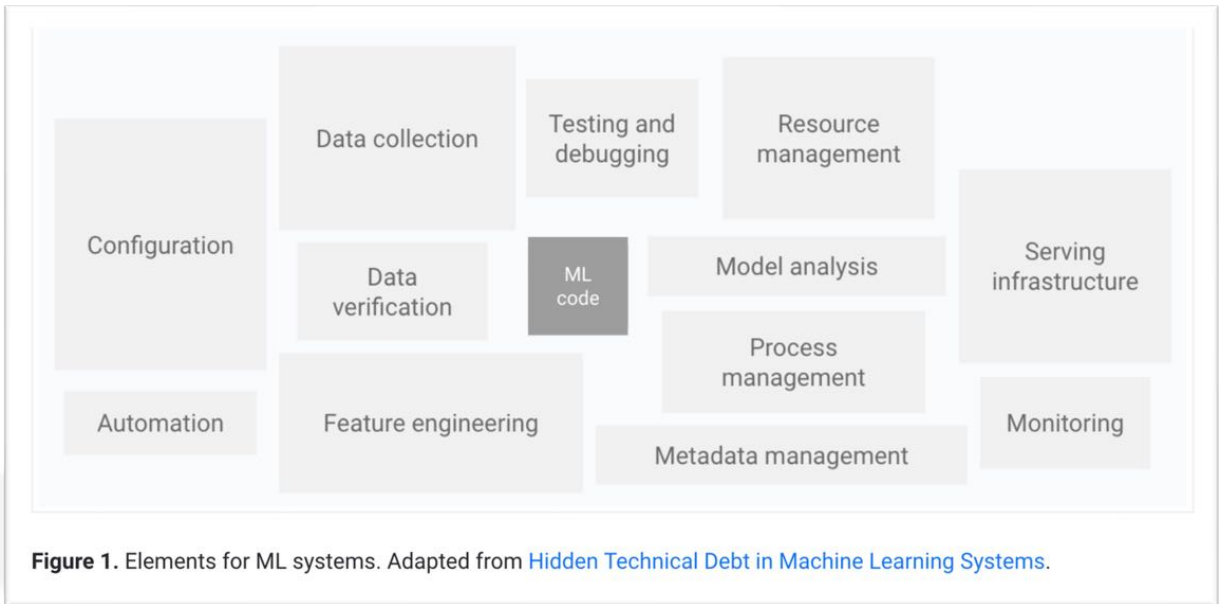


Figure 1. Elements for ML systems. Adapted from Hidden Technical Debt in Machine Learning Systems.

Reflection.

From our examples, the code (‘algorithm’) is only a small part of it!

Transparency is required in planning, implementation, auditing…

… concerns the data, design, <actual ML stuff here>, testing, deployment.

See e.g. Pang and Lee (2004) in Gebru et al (2021) https://arxiv.org/pdf/1803.09010.pdf -->

… also consider legal aspects & philosophical concepts in the broader sense: incl. fairness, recourse…

Movie Review Polarity — Thumbs Up? Sentiment Classification using Machine Learning Techniques

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the exception that no more than 40 posts by a single author were included (see “Collection Process” below). No tests were run to determine representativeness.

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example “negative polarity” instance, taken from the file neg/cv452_tok-18656.txt.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in “Data Preprocessing”).

Is there a label or target associated with each instance? If so, please provide a description.

The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information (such as threads, replies, or posts by the same author) could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The instances come with a “cross-validation tag” to enable replication of cross-validation experiments; results are measured in classification accuracy.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

See preprocessing below.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links

[1] All information in this datasheet is taken from one of the following five sources; any errors that were introduced are the fault of the authors of the datasheet: http://www.cs.cornell.edu/people/pabo/movie-review-data/, http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

Fig. 1. Example datasheet for Pang and Lee’s polarity dataset [22], page 1.

Current issues in transparency:

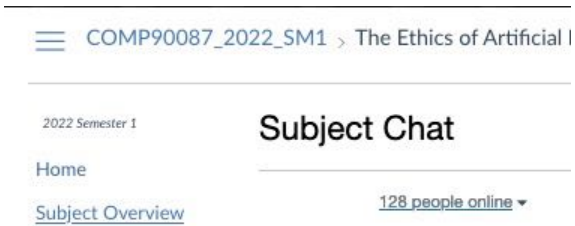
Social media advertising systems

Image by DALLE2

Audience activity IV [2-5 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.

Incentive:

Image source: Cadbury

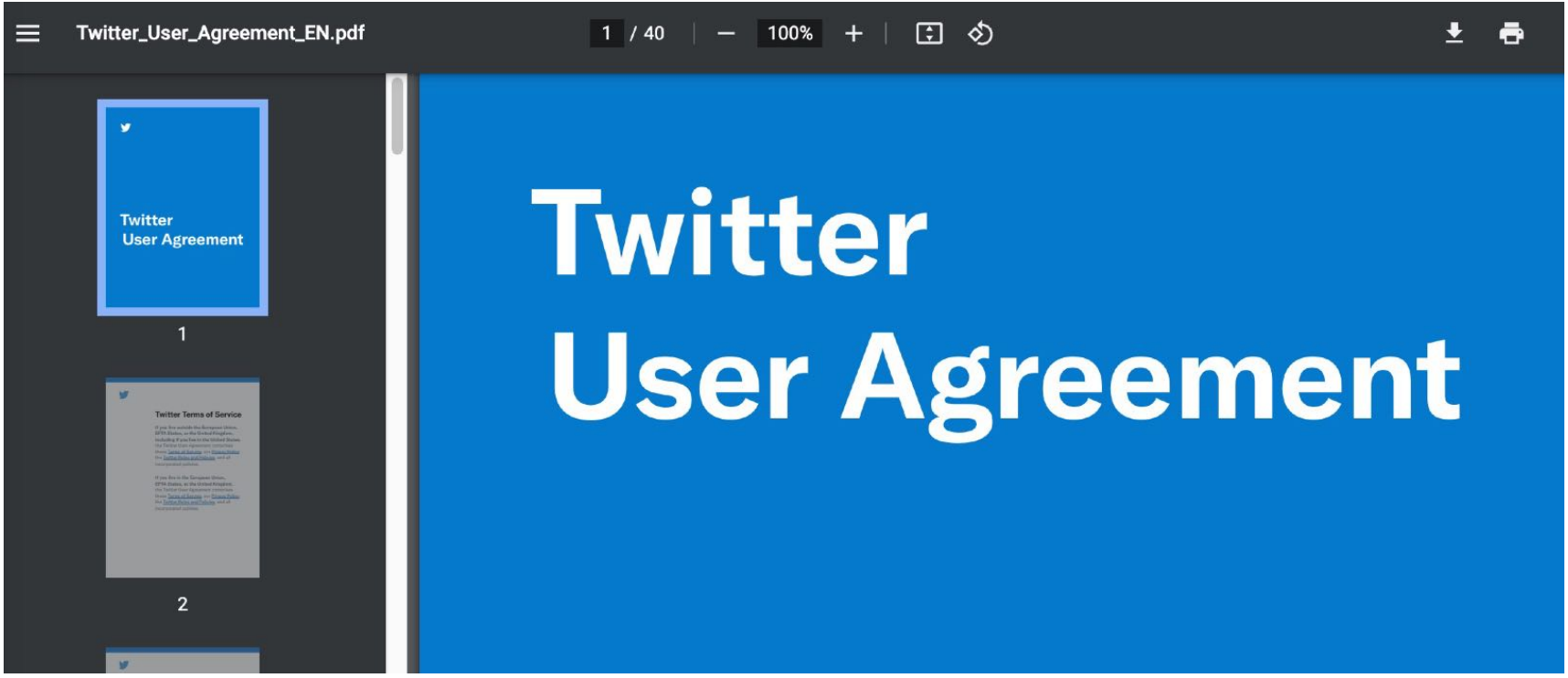
When printed out, can you guess how long Facebook or Twitter's policies and terms of service are.

Let's say default font, on A4 paper?

The big question: do you read them?

Image source: twitter.com/en/tos

Consider Twitter



TL;DR: Twitter?

Summarised by TOSDR

(Source: https://tosdr.org/en/service/195)

Even the key points take up ~3 A4 pages;
refer image →

Consider Meta's Facebook

Summarised by TOSDR

(Source: https://tosdr.org/en/service/182)

Even the key points take up ~3 A4 pages →

Roughly 10 pages in A4 printed (as of 2021).

This doesn't even include the additional policies:
e.g. 'Commercial Standards', 'Advertising Policies, …
(~14 more links)

☕ Break time!

See you in 5 mins.

Image by DALLE2

Image source: Meta

Consider ads: is this 'transparent'?

Let's say I've been shown an ad (sponsored post) on Facebook. As a consumer, I want to know WHY.

I go to the TOS, then find 'Ads', then find 'Data Policy'…

Which takes me to another page…

Which scares me.

Let's try another method – go to the ad, and click on the menu.

Device Information

As described below, we collect information from and about the computers, phones, connected TVs and other web-connected devices you use that integrate with our Products, and we combine this information across different devices you use. For example, we use information collected about your use of our Products on your phone to better personalize the content (including ads) or features you see when you use our Products on another device, such as your laptop or tablet, or to measure whether you took an action in response to an ad we showed you on your phone on a different device.

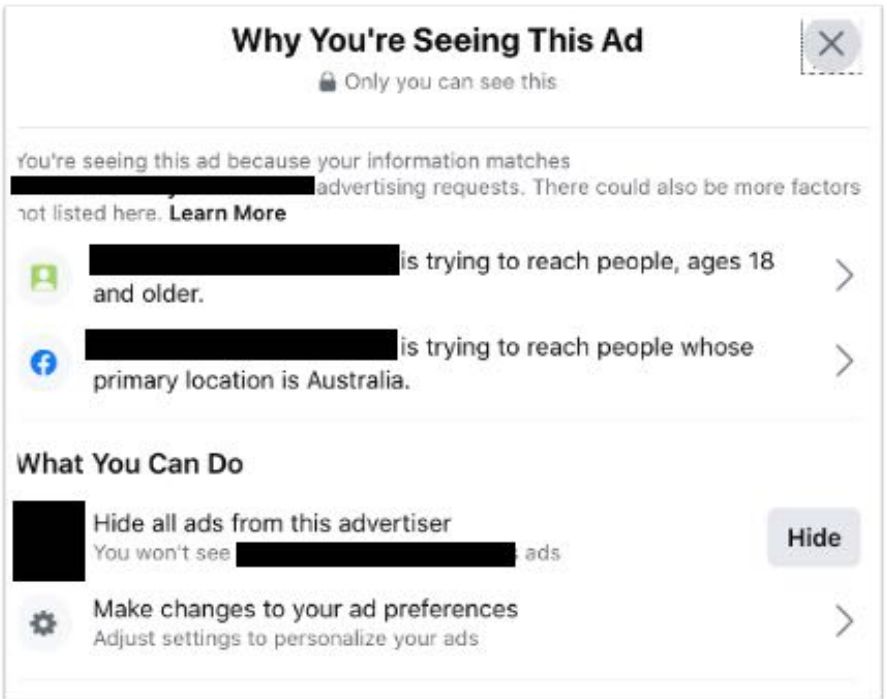
Information we obtain from these devices includes:

- **Device attributes:** information such as the operating system, hardware and software versions, battery level, signal strength, available storage space, browser type, app and file names and types, and plugins.
- **Device operations:** information about operations and behaviors performed on the device, such as whether a window is foregrounded or backgrounded, or mouse movements (which can help distinguish humans from bots).
- **Identifiers:** unique identifiers, device IDs, and other identifiers, such as from games, apps or accounts you use, and Family Device IDs (or other identifiers unique to Facebook Company Products associated with the same device or account).
- **Device signals:** Bluetooth signals, and information about nearby Wi-Fi access points, beacons, and cell towers.
- **Data from device settings:** information you allow us to receive through device settings you turn on, such as access to your GPS location, camera or photos.

Image source: Meta

Consider ads: is this 'transparent'?

Let's say I've been shown an ad (sponsored post) on Facebook. As a consumer, I want to know WHY.
(Image sources: Facebook)



Location information:

We may use your location information to show you ads from advertisers trying to reach people in or near a specific place. We get this information from things such as:

- Where you connect to the internet.
- Where you use your phone.
- Your location from your Facebook and Instagram profile.

Consider ads: is this 'transparent'?

Case study thanks to Michael Geers, (Max Planck Institute for Human Development, Germany)

Lorenz-Spreen, P., Geers, M., Pachur, T., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2020). *A simple self-reflection intervention boosts the detection of microtargeted advertising*. https://doi.org/10.31234/osf.io/ea28z

Let's say I've been shown an ad (sponsored post) on Facebook.
As a consumer, I want to know WHY.

FB gives me some reasons.

… but also "more factors not listed".

… and directs me to a huge page with many explanations, but all "mays" (we may, advertisers may…)

… and has a long TOS explaining the many ways.

Consider ads: is this 'transparent'?

Case study thanks to Michael Geers, (Max Planck Institute for Human Development, Germany)

Lorenz-Spreen et al (2021)

P4-5:
"At present, the platforms' transparency measures offer **"nominal transparency",** with no real regard for whether people actually can easily access, read and gain insight into the information held about them and whether this transparency in name foster users' autonomy.

"Aiming for effective transparency—which demonstrably enables users to understand what platforms do with their data and what users' choices imply, and to then translate this knowledge into measurable behaviour—is an important step towards more acceptable business practices and towards regaining some of the lost autonomy for users (e.g., by prompting people to adjust their privacy settings; Parra-Arnau et al., 2017)

Current issues in transparency:

Criminal justice AI systems

Image by DALLE2

Criminal Justice and AI systems

You have seen some of the following examples in your exploration of AI ethics. (Images are from Wikipedia)

COMPAS (software)

From Wikipedia, the free encyclopedia

Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.[1][2]

COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions.[3]

PredPol

From Wikipedia, the free encyclopedia

PredPol, Inc is a predictive policing company that attempts to predict property crimes using predictive analytics. PredPol is also the name of the software the company produces. PredPol began as a project of the Los Angeles Police Department (LAPD) and UCLA professor Jeff Brantingham. PredPol has produced a patented algorithm, which is based on a model used to predict earthquake aftershocks.

PredPol

Type	Private
Headquarters	Santa Cruz
Products	Predictive analytics
Website	www.predpol.com

As of 2020, PredPol's algorithm is the most commonly used predictive policing algorithm in the U.S.[1][2] Police departments that use PredPol are given printouts of jurisdiction maps that denote areas where crime has been predicted to occur throughout the day.[3] The *Los Angeles Times* reported that officers are expected to patrol these areas during their shifts, as the system tracks their movements via the GPS in their patrol cars.[4] Scholar Ruha Benjamin called PredPol a "crime production algorithm," as police officers then more heavily patrol these predicted crime zones, expecting to see crime, which leads to a self-fulfilling prophecy.[1]

Reading: Rudin (2019) on COMPAS

The focus of this case study is not about the technical explainability for the underlying algorithms, etc.

… but about the transparency of the processes and decisions involved.

Take COMPAS – and its **decision making** assumptions, in practice (Rudin 2019)

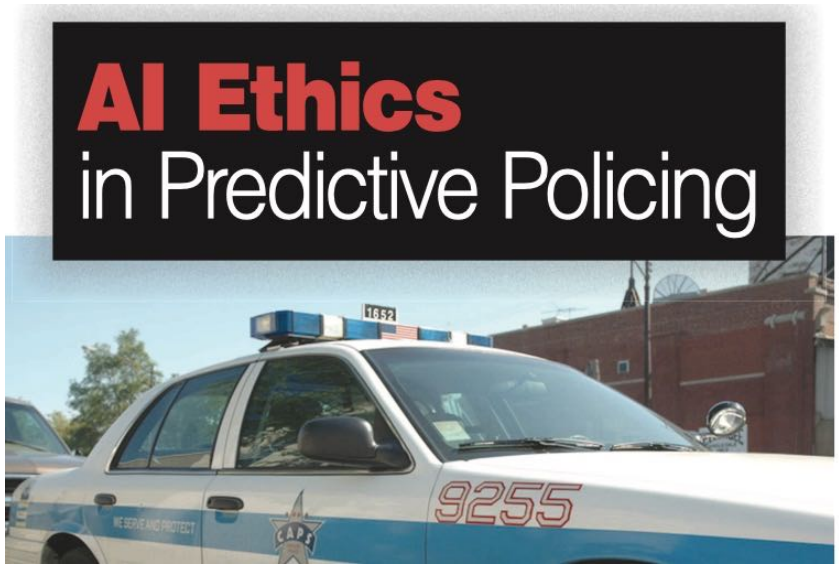
Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

Cynthia Rudin
Duke University

But if the model is a black box, it is very difficult to manually calibrate how much this additional information should raise or lower the estimated risk. This issue arises constantly; for instance, the proprietary COMPAS model used in the U.S. Justice System for recidivism risk prediction does not depend on the seriousness of the current crime [27, 29]. Instead, the judge is instructed to somehow manually combine current crime with COMPAS. Actually, it is possible that many judges do not know this fact. If the model were transparent, the judge could see directly that the seriousness of the current crime is not being considered in the risk assessment.

Asaro (2019) on PredPol

In the same vein we consider COMPAS;
cf PredPol in Asaro (2019)



Transparency over the algorithms, data, and practices of implementation are also necessary. While the Chicago Police Department sought to avoid embarrassment from releasing the details of the SSL, it would be impossible for independent outside researchers to evaluate its impacts — positive and negative — without access to the data and algorithms. It should not take a prolonged lawsuit from a newspaper for government agencies to share public data. Of course, as more and more commercial systems, like PredPol,[19] make the algorithms and even the data proprietary, they will fall under intellectual property protections. This means private companies will be processing the data, and will not be required to reveal their algorithms, or subject them to independent outside scrutiny. In some cases, private

Ferguson (2017) on PredPol

In the same vein we consider COMPAS; cf. PredPol - Ferguson (2017) provides a legal perspective.

Issues at all stages: crime stats; data dossiers; personal/cultural bias; data entry/analysis; tech complexity; financial/IP interests; auditing; metrics…

Washington University
Law Review

VOLUME 94 NUMBER 5 2017

POLICING PREDICTIVE POLICING

ANDREW GUTHRIE FERGUSON*

1. Transparency: Vulnerabilities

As currently implemented, a lack of transparency exists at all levels of predictive policing. Even something as simple as crime statistics, which in many cases are publicly available, remain rife with concerns about accuracy and completeness.[316] Adding personal data dossiers to these crime statistics creates new problems, as the sheer volume of information complicates transparent assessment of the sources underlying the predictions.[317] How do you fix an error in the data if you cannot see that such an error exists? How do you even know who has the responsibility to input information into these big aggregated databases?[318] In addition, unintended personal or cultural biases can infect the data, the scoring systems, the source codes, and thus the resulting predictive outcome.[319] Simply stated, without significant investment in exposing the data collection methods, weaknesses, and gaps, and without equal investment in understanding the challenges associated with inputting and analyzing the data, the entire system runs the risk of being built on an unknown and unknowable database.[320]

The nature of algorithms further obscures the process, except perhaps to technical experts. Police officers and administrators receive the results, but due to the complexity of the chosen algorithm they can rarely understand the underlying math. Thus, predictive policing runs into the same problems as other automated predictive technologies: the technical complexity of the design makes it nearly impossible for outsiders to determine the accuracy, effectiveness, or fairness of the program.[321] True, police can see if the system works, but police cannot see *how* the system works. This lack of transparency is not simply the result of new technology, but also the influence of the proprietary nature of the software. The companies involved in these real-world tests are in a multimillion-dollar race to convince police departments to adopt their particular products. The companies have financial interests and proprietary secrets to protect, and every incentive to report positive outcomes.[322]

Effectiveness itself remains a contested issue. Early tests show a correlation between use of certain predictive policing techniques and decreased crime rates (for some crimes). But how do police districts determine metrics in the future? Crime may go up or down independent of the chosen computer program. Crime analysts may make a more or less accurate comparative judgment. Most importantly, how can outsiders audit the data? In similar police data collection experiments (DNA databases, "stop and frisk" reporting), the police have audited themselves with mixed results.[323]

Reflection.

AI systems in justice/policing causes serious effects on people's freedom and status under the law.

The legal perspective of AI ethics gives us another perspective on the need for transparency.

"Transparency is difficult, but it matters to a functioning predictive system that deals with individuals' lives and liberty" (Ferguson, 2017).

How do we start fixing the issues?
E.g. for PredPol - auditing, public release of metrics, training (Ferguson, 2017).

…or not use PredPol to begin with?

Current issues in transparency:

Generative AI!

Image by DALLE2

Issues in Generative AI

One key issue is the lack of transparency in the data sets used to train generative AI systems. These systems are often trained on large data sets, which can contain biases and other issues that can affect the output of the system. Without transparency in the data sets, it can be difficult to identify and address these issues.

Another issue is the lack of transparency in the algorithms used in generative AI systems. (Re: Tim's lecture)

A third issue is the lack of transparency in the decision-making processes used by generative AI systems. These systems can make decisions based on complex calculations and analysis, but it can be difficult to understand how these decisions are being made. This can be particularly concerning in cases where the system is making decisions that can have a significant impact on people's lives, such as in medical diagnosis or hiring decisions.

Image source: KnowYourMeme

Issues in Generative AI

One key issue is the
are often trained on
the system. Without

Another issue is the

A third issue is the la
These systems can m
understand how thes
system is making dec
or hiring decisions.

ain generative AI syste
other issues that can a
t to identify and address these ues.

generative AI systems. (Re: Tim's lecture)

cesses used by generative AI systems.
nd analysis, but it can be difficult to
cularly concerning in cases where the
ople's lives, such as in medical diagnosis

I only wrote this

Issues in Generative AI

Let's get ChatGPT (GPT-3.5 🤖) to set the outline…

Fair warning: there are things it doesn't get right, so we have to supplement this with our own discussions! Hence, which is why we say that ChatGPT is still not good at philosophical argumentation.

🤖 "lack of transparency in the data sets used to train generative AI systems" (ChatGPT, 2023).

🤖 "lack of transparency in the decision-making processes used by generative AI systems" (ChatGPT, 2023).

- "particularly concerning in cases where the system is making decisions that can have a significant impact on people's lives, such as in medical diagnosis or hiring decisions."

👨‍🏫 **Lack of transparency on the effects on humans/society.**

Lack of transparency in the data sets?

🤖 “lack of transparency in the data sets used to train generative AI systems” (ChatGPT, 2023).

A “detector would be built into ChatGPT to check whether it was echoing the toxicity of its training data, and filter it out before it ever reached the user...

To get those labels, OpenAI sent tens of thousands of snippets of text to an outsourcing firm in Kenya, beginning in November 2021. Much of that text appeared to have been pulled from the darkest recesses of the internet.” → **traumatising**!

TIME

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

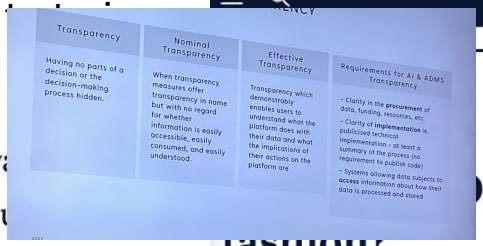
Perrigo, B. (2023, January 18). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Lack of transparency in the data sets?

🤖 “lack of transparency in the data sets used [illegible] generative AI systems” (ChatGPT, 2023).

Plagiarism must also be considered, Cheong says. If the softw[illegible] existing content from the web, could copyright become an issu[illegible] the algorithm learns from discriminatory material?

“Imagine an up-and-coming artist with exceptional memory and attention to detail set off to learn the works of thousands of artists worldwide, learning to create new images in the style of any artist ‘by request’ – from van Gogh to an amateur painter – with little effort,” says Cheong. “Any artist who might have had their work viewed by this new maestro will be concerned that the same could happen to their art.”



Geraets, N. (2022, November 11). Dress code: Does AI technology belong in fashion? *The Sydney Morning Herald*. https://www.smh.com.au/business/companies/dress-code-does-ai-technology-belong-in-fashion-20221108-p5bwh2.html

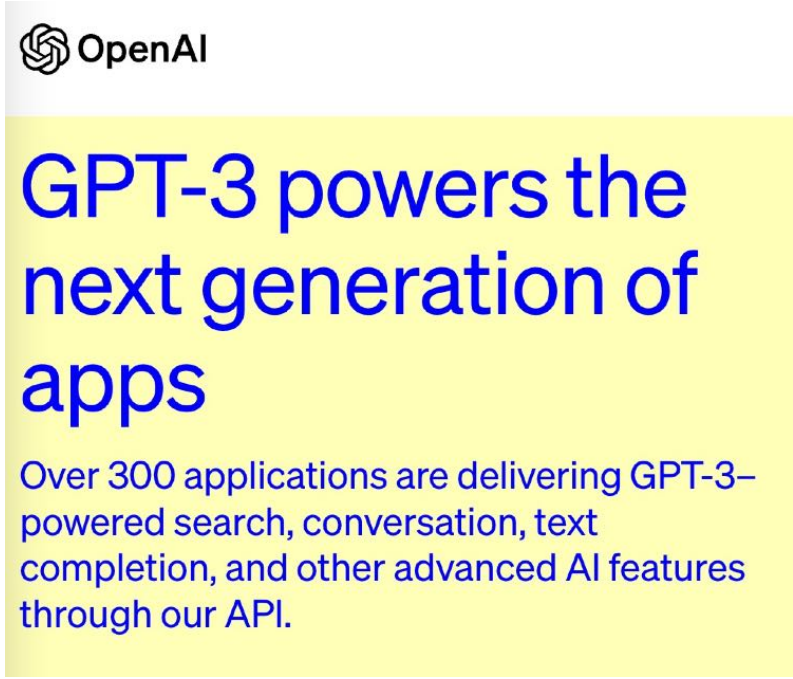
Image sources: https://openai.com/blog/gpt-3-apps and MS Office clipart.

Lack of transparency in decision making?

🤖 "lack of transparency in the decision-making processes used by generative AI systems" (ChatGPT, 2023).

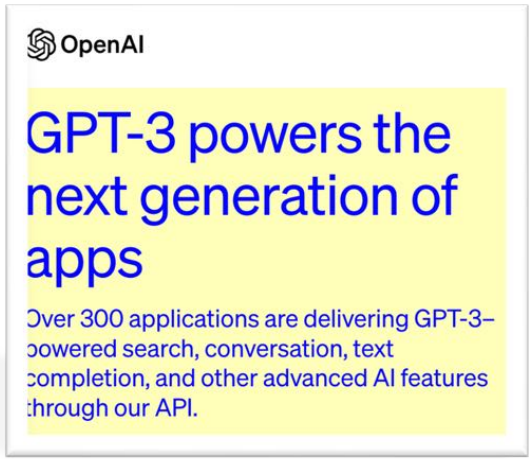
Straightforward case: "medical diagnosis"

Not so straightforward: Downstream use!



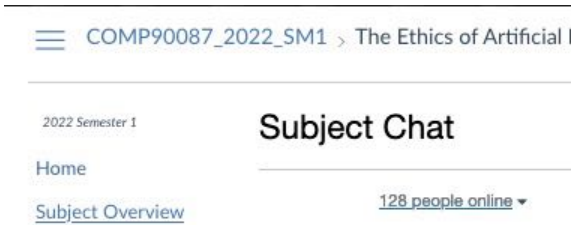
Audience activity V [2-5 mins]

Image source: https://openai.com/blog/gpt-3-apps



Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.

Incentive:

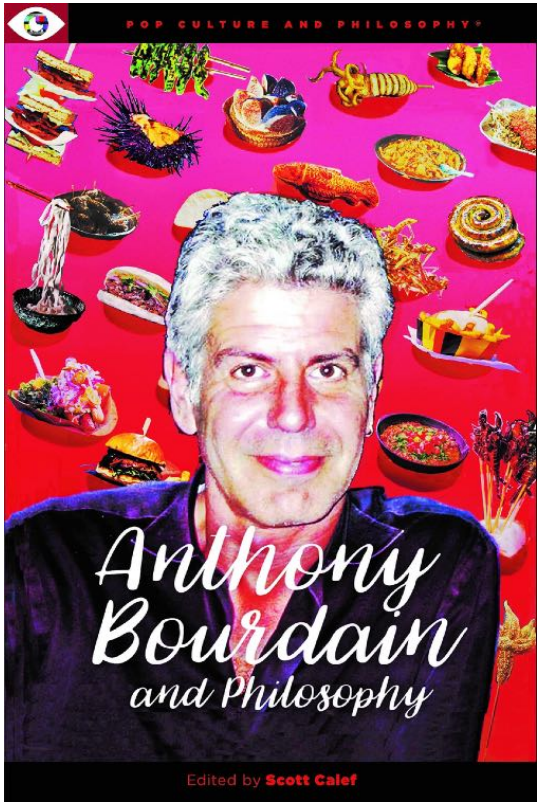
Image source: Cadbury

What are the DOWNSTREAM implications on (potential lack of) transparency in decision making based on Generative AIs?

Discuss with reference to the 300+ apps (and counting) that rely on GPT-3 (and the many more based on Stable Diffusion, etc).

Image source: Amazon/Carus Books

Lack of transparency: human/societal impact?



"After the untimely death of Anthony Bourdain, director Morgan Neville sought to narrate his life in the documentary Roadrunner (by Focus Features, 2021). …

Several snippets of speech in Roadrunner which sounded like Anthony were originally written by him, but, contrary to first impressions, were not actually vocalized by Bourdain. The moviemakers used an Artificial Intelligence (AI) 'doppelgänger' to simulate Bourdain's voice reading out sentences he never actually narrated in real life."

(Cheong, 2023 –
"Chef of the Future?" The ethics of 'deepfaking' Bourdain after death;
in Anthony Bourdain and Philosophy, edited by Scott Calef).

Image source: NYT / Jason Allen

Lack of transparency: human/societal impact?

Image source: NYT / Jason Allen

Lack of transparency: human/societal impact?

https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

By Kevin Roose

Sept. 2, 2022

This year, the Colorado State Fair's annual art competition gave out prizes in all the usual categories: painting, quilting, sculpture.

But one entrant, Jason M. Allen of Pueblo West, Colo., didn't make his entry with a brush or a lump of clay. He created it with Midjourney, an artificial intelligence program that turns lines of text into hyper-realistic graphics.

Mr. Allen's work, "Théâtre D'opéra Spatial," took home the blue ribbon in the fair's contest for emerging digital artists — making it one of the first A.I.-generated pieces to win such a prize, and setting off a fierce backlash from artists who accused him of, essentially, cheating.

Big Data Research & Social Media: *From Elections to Pandemics*

Image by DALLE2

Reading: Walsh (2019)

Experiments in Social Media

Author: Toby Walsh[1]

Abstract.

Social media platforms like Facebook and Twitter permit experiments to be performed at minimal cost on populations of a size that scientists might previously have dreamt about. For instance, one experiment on Facebook involved over 60 million subjects. Such large scale experiments introduce new challenges as even small effects when multiplied by a large population can have a significant impact.

Recent revelations about the use of social media to manipulate voting behaviour compound such concerns. It is believed that the psychometric data used by Cambridge Analytica to target US voters was collected by Dr Aleksandr Kogan from Cambridge University using a personality quiz on Facebook. There is a real risk that researchers wanting to collect data and run experiments on social media platforms in the future will face a public backlash that hinders such studies from being conducted. We suggest that stronger safe guards are put in place to help prevent this, and ensure the public retain confidence in scientists using social media for behavioural and other studies.

Not just Cambridge Analytica – which was for election targeting etc.

Walsh (2019) found that studies by academics "to improve voter participation" in fact "increased turnout by about 340,000 additional votes" (citing Bond et al, 2012).

Issue (A) – election manipulation?

Reading: Walsh (2019)

> The first recommendation is that we may need to take into account not just the impact on the individual under study but the broader impact any experiment might have on society. For a study on voting, this might be an electoral risk. For a study on fake news, it might be decreasing trust within society in real news. For a study on manipulating people's emotions, it might be the emotional wellbeing of the population studied.

Reflections for transparency in social media and big data research, by Walsh (2019):

1. "The first recommendation is that we may need to take into account not just the impact on the individual under study but the broader impact any experiment might have on society…"

2. "The second recommendation is that ethics approval may be needed…"

3. "The third recommendation is that subjects of any experiment may need to be informed directly after the study about the results and their participation..."

Issue (B) – Contemporary issue in research re: #3?

Conclusion.

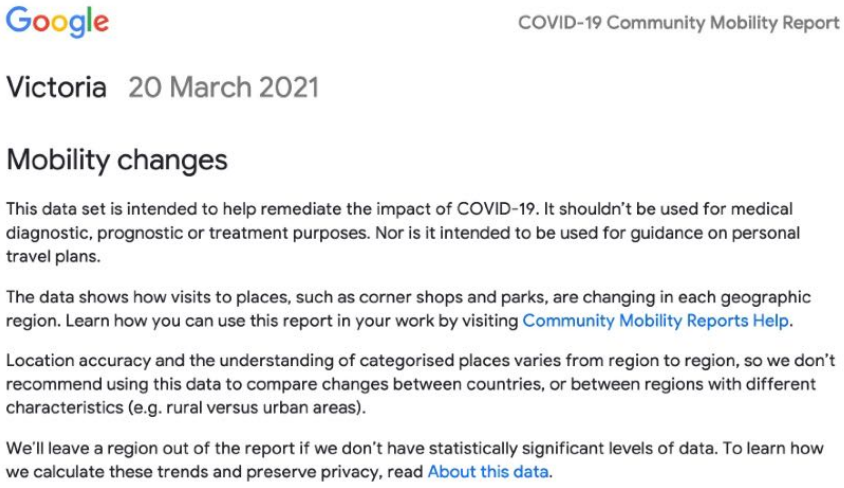
Big data research on social media invokes many concerns – privacy (can the user opt out of the 'researchers gaze'); autonomy (does the research make the users do things they won't otherwise?); wellbeing (does the research have the potential to change mood/health outcomes?)… **are these clear to the users?**

Concluding food for thought: is this ethical?

How Facebook and Google Track Public's Movement in Effort to Fight COVID-19

Location data provide rich resource for decision makers, scientists, and the public

By **Emily Waltz**



Google

COVID-19 Community Mobility Report

Victoria 20 March 2021

Mobility changes

This data set is intended to help remediate the impact of COVID-19. It shouldn't be used for medical diagnostic, prognostic or treatment purposes. Nor is it intended to be used for guidance on personal travel plans.

The data shows how visits to places, such as corner shops and parks, are changing in each geographic region. Learn how you can use this report in your work by visiting Community Mobility Reports Help.

Location accuracy and the understanding of categorised places varies from region to region, so we don't recommend using this data to compare changes between countries, or between regions with different characteristics (e.g. rural versus urban areas).

We'll leave a region out of the report if we don't have statistically significant levels of data. To learn how we calculate these trends and preserve privacy, read About this data.

Sources: https://spectrum.ieee.org/the-human-os/telecom/wireless/facebook-google-data-publics-movement-covid19; Google

Any concluding discussions?

(Over to Prof Jeannie Paterson for Week 7…

…and I'll be back in a few weeks)

Image by DALLE2

THE UNIVERSITY OF MELBOURNE

Thank y