LA TROBE UNIVERSITY

Building AI – Module 6

AI Ethics

21 June 2022

Optus U

Agenda

1. Defining ethics
2. Data vs AI ethics
3. Ethics frameworks
4. Ethics in practice
5. Public interest technology

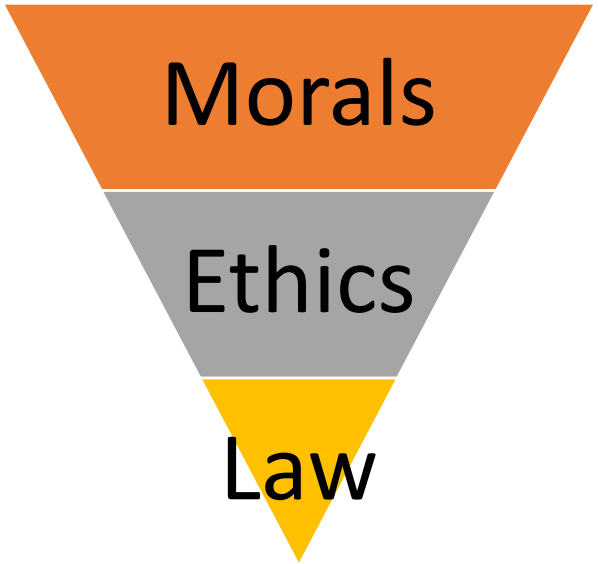
Optus U

Why AI Ethics?

- AI enabled/associated technologies are becoming an integral part of human society
  - Contrast this with archaic technology – digital calculator, wristwatch, CD player
- Acquiring “data” – strict regulations and standards, also inevitable
- Acquiring an “intelligence” from data – not so much
  - Not as severe as a data breach
  - Not as slight as employee access
- AI Ethics equally applies to ‘data analytics’ and ‘data science’

Moral, ethical, legal

- Moral, ethical, legal
  - Morals – individual, personal beliefs, philosophical
  - Ethics – community, correct behaviour, analytical
  - Law – society/country, a basic standard that is enforced to protect morals and ethics, logical



Defining Ethics

- Many formal definitions (‘the science of the ideal human character’)
- A branch of philosophy that can be summarised as “What should I do?” or “How should I live?”
- A critical phase in any decision-making process guided by values, principles and purpose, instead of own interests, social convention, company policy, or even rule of law.
  - Values – what is good, what we strive, desire and seek to protect.
  - Principles – what is right, in terms of achieving our values (above).
  - Purpose – in life, rationalises values and principles

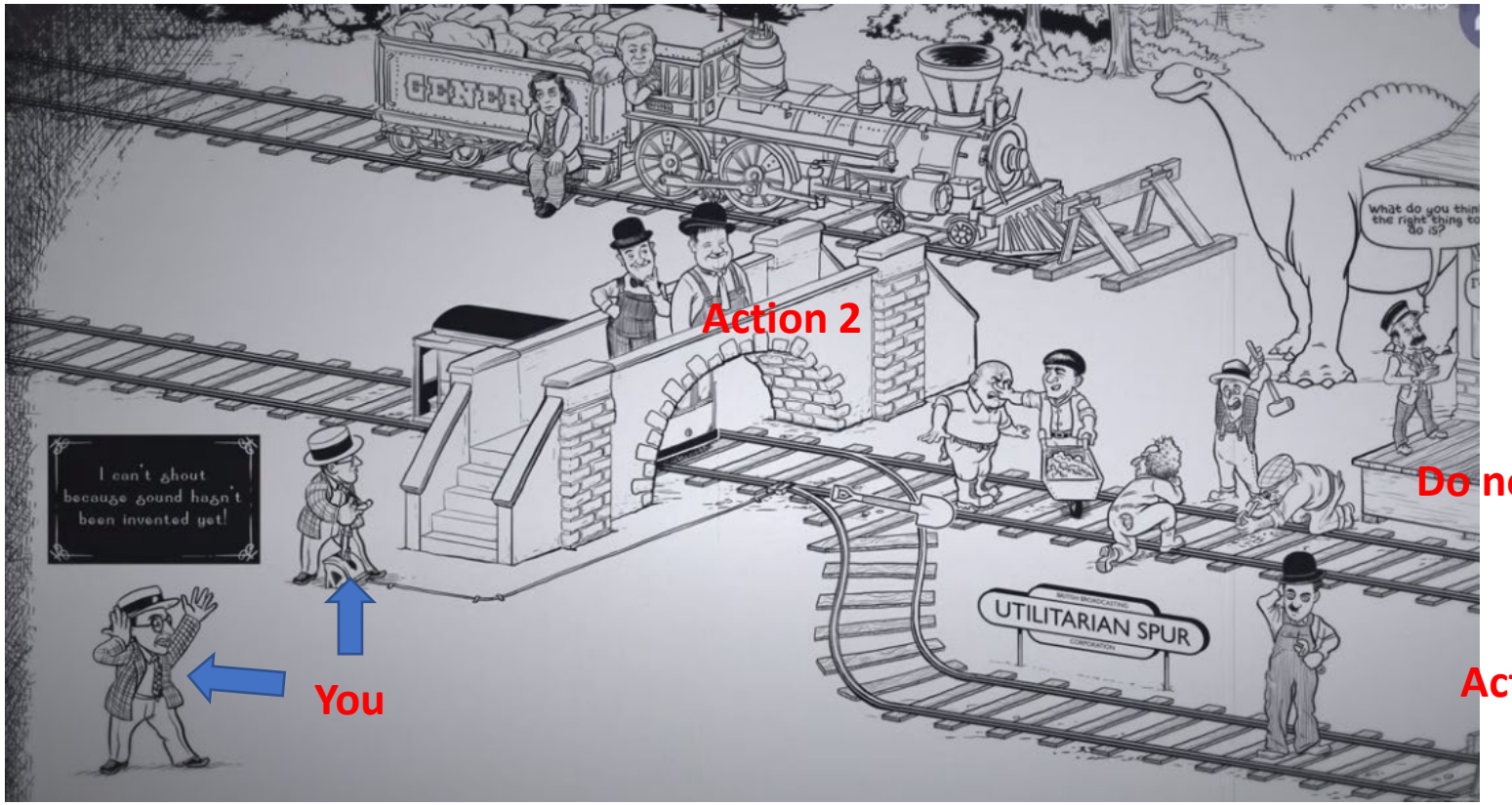
Ethics – Areas of Study

- Meta-ethics – how we understand ethics (theory and philosophy)
- Normative ethics - study of ethical action, standards for right/wrong
- Applied ethics – application to real-world settings
  - Medical ethics
  - Bioethics
  - AI ethics
  - Animal ethics
  - Business ethics etc.

Ethics theories

- Consequentialist – an action is judged by its consequences
  - Utilitarianism – ‘greatest good for the greatest number’
  - Ethical egoism/altruism - self-centred vs selfless
- Non-consequentialist – an action is judged on properties intrinsic to the action
  - Aristotle’s virtue ethics – a character based approach of virtue through practice
  - Kant’s deontological (duty-based) ethics - rules to distinguish right from wrong
- The Dark Knight
  - Batman should kill the Joker
  - Batman should not kill the Joker
  - What kind of person would kill the Joker?

Trolley Dilemma



https://www.youtube.com/watch?v=bOpf6KcWYyw

Similar - COVID vaccines and rare blood clotting

AI Ethics in the public interest

- Autonomous cars
  - Networks of movement, convenient, economical, efficient
  - Avoiding a road accident, saving a human life, which human?
- Automated sentencing
  - Big data-driven insights that overcome human bias, fast, informed
  - Inclusivity, completeness, quality – human bias is easily replicated in training datasets
- Personalisation and disinformation overload
  - Social beings, networks of support, emergency announcements
  - Confirmation bias, social herding, echo chambers, fake news bots
- Military and surveillance use cases
- Those who build the AI – 80% of AI academics and 70% of AI industry are male

A Current Debate

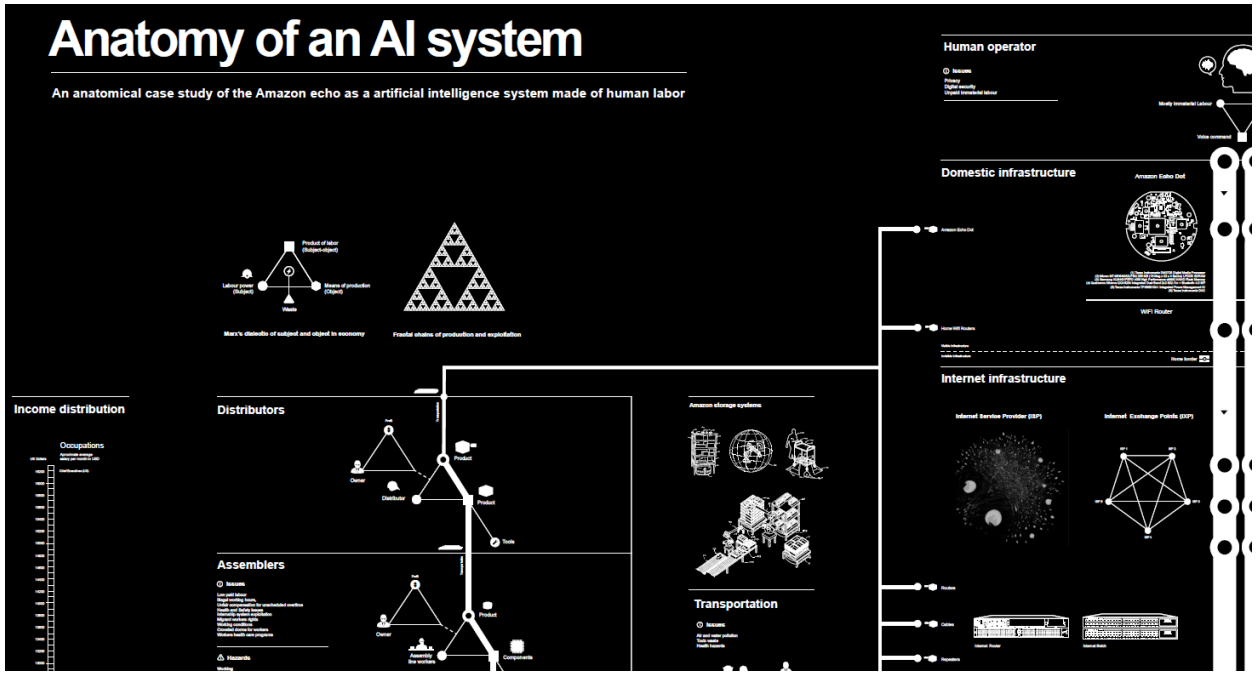
Geoffrey Hinton
@geoffreyhinton

Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

7:37 AM · Feb 21, 2020 · Twitter Web App

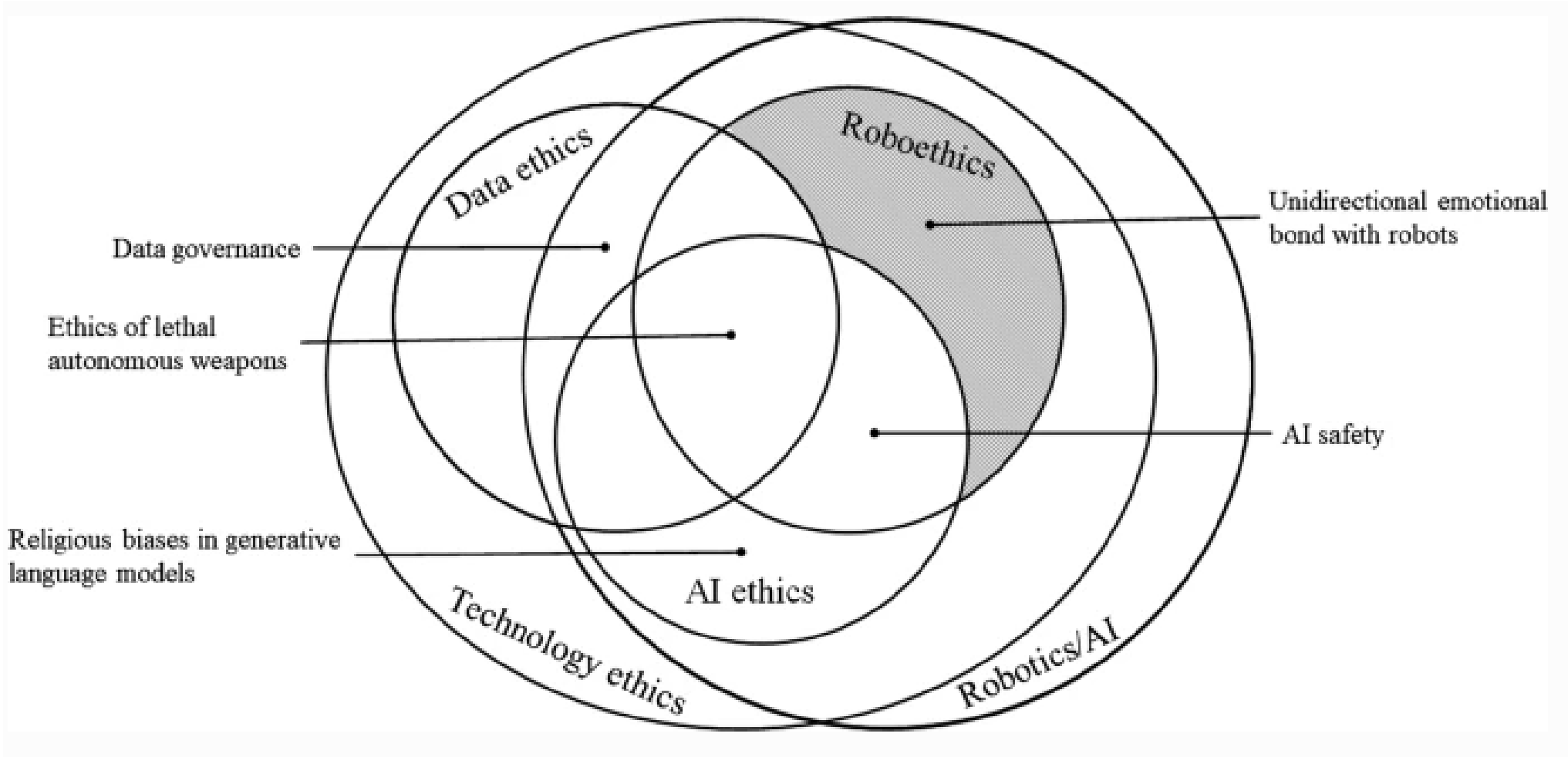
1,158 Retweets **614** Quote Tweets **5,175** Likes

https://twitter.com/geoffreyhinton/status/1230592238490615816



https://anatomyof.ai/img/ai-anatomy-map.pdf

Which AI ethics?



Data - Security, Protection, Privacy

- Data security – securing data from external threats (unauthorised access to systems)
  - Confidentiality - only accessed by authorised individuals
  - Integrity - not altered without authorisation
  - Availability - reliable access to data
- Data protection - securing data from internal threats (unregulated control of data)
  - A legal construct for informational self-determination, data loss
- Data privacy – how the data gets used (unspecified processing of data)
  - Consent and notice
  - Legal right to access/delete
  - Third party access

GDPR

- General Data Protection Regulation
  - A law on data protection and privacy in the European Union and the European Economic Area.
  - The strongest set of data protection rules in the world, implemented in May 2018
  - Gives control of personal data and simplify data access for organisations
- Seven principles:
  - Lawfulness, fairness and transparency
  - Purpose limitation
  - Data minimisation
  - Accuracy
  - Storage limitation
  - Integrity and confidentiality (security)
  - Accountability

Article 5(1) requires that personal data shall be:

> "(a) processed lawfully, fairly and in a transparent manner in relation to individuals ('lawfulness, fairness and transparency');
>
> (b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall not be considered to be incompatible with the initial purposes ('purpose limitation');
>
> (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');
>
> (d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy');
>
> (e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes subject to implementation of the appropriate technical and organisational measures required by the GDPR in order to safeguard the rights and freedoms of individuals ('storage limitation');
>
> (f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality')."

Article 5(2) adds that:

> "The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('accountability')."

In Australia..

- Most states and territories (except WA and SA) have their own data protection legislation applicable to state government agencies, and private businesses that interact with state government agencies.
- At the Federal level:
  - Privacy act 1988: https://oaic.gov.au/privacy-law/
  - Australian Privacy Principles
  - Privacy Amendment (Notifiable Data Breaches) Act 2017: https://www.oaic.gov.au/privacy-law/privacy-act/notifiable-data-breaches-scheme
  - Who – government agency, business with +3m turnover, organisations in health services
- Australian Government Information Security Manual
  - https://www.cyber.gov.au/ism

Consumer Data Right

- CDR establishes the right for consumers to direct a supplier to share designated data about the consumer with another supplier or with the consumer themselves
- CDR aims to increase the bargaining power of consumers by enabling comparison and switching between products/services, while also driving competition and innovation between service providers
- Introduced into the banking sector in July 2020 (open banking), energy and telecom to follow
- It is an opt-in system, the consumer can authorise a business to access your data, with control over what data is transferred, where it is used, who it can be disclosed to, with the ability to withdraw consent at any time and delete if no longer needed
- Only ‘accredited data recipients’ accredited by the ACCC can provide services, which include requesting data with the customer’s consent, then using to provide a comparison or product/service

Responsible data collection

- Two rules of thumb – avoid harm and build trust
- Manipulation and misrepresentation
  - Manipulate consumers into revealing personal information (apps).
  - Misrepresented identity (collect survey data in the guise of student projects).
- Develop an awareness and accountability of the following ten principles
  - Informed consent – should the subject know data is being collected and agree to its collection?
  - Anonymity - should all personally identifying information be eliminated from the data? or collect only in the form of aggregates such that individuals cannot be identified?
  - Confidentiality - should sources and providers of data be protected from disclosure?
  - Security - what level of protection from intrusion, corruption, and unauthorized access?
  - Privacy - should each individual have the ability to control access to personal data about themselves?

Ten principles (continued)

- Accuracy - what level of exactness and correctness is required of the data?
- Ownership - is personal data about individuals an asset that belongs to the business or privately owned information for which the business has stewardship responsibilities?
- Honesty - to what degree should the business be forthright and visible about data collection practices?
- Responsibility - who is accountable and at what level for use and misuse of data?
- Transparency – between the two extremes of open and stealth data collection, what is the right level of transparency?

Anonymised data

- Anonymised data is frequently used in AI projects.
- Remove/replace classification of personally identifiable information so that individuals associated with that data can remain anonymous.
- Useful for segmentation - identifying collective patterns which does not require information at individual level.
- But, the more Anonymised the less useful it becomes.
- Identity information inevitably removes contextual information.
- And pseudo-anonymised
- A pragmatic solution – be transparent and provide consumers the choice to opt-in/opt-out.

Opt-in and Opt-out

- Seen on many websites and social networks.
- Opt-out – default settings used by the service provider most often for all data collection.
  - The user must explicitly choose to opt out of the default into custom settings.
- Opt-in – explicit permission has to be granted to collect and use information in a certain set of ways before the collection of data begins.
  - Tedious but useful as it forces end-users to consider repercussions before making a choice.
  - Less likely to be used as it's frequently ignored (need to incentivise).

Data - other considerations

- The right to be forgotten – explicit request for removal of data.
- Individuals requesting invisibility from search engines raises questions about both privacy and freedom of speech.
- The right to data expiry - data be deleted after it fulfills a business purpose
- Data ownership – provide the end-user ownership/full control of their data
- Lessons from bioethics - medical/clinical research

On to AI ethics

- Technology ethics, automation ethics, AI ethics, robot ethics, machine ethics, etc…
- Unlike legally binding regulations for data (like the GDPR), AI ethics guidelines and frameworks are non-legislative policy instruments
- “The use of technology (and intelligence) in tasks and functions that are generally repetitive to achieve an equal or better performance than the human experts and operators who perform the same” (Parasuraman, 1997)
- Autonomous and intelligent technical systems that decrease the need for human intervention in organisations, society, and industries can also have a negative impact on individuals and societies.
- Because of this disposition, these technologies must be “aligned” to values and principles upheld by societies
- This “ethics collective” aims to enable and achieve this alignment.

Rationale for AI Ethics

- Diversity of data
- Non-linearity of algorithms
- Impact of insights
- Consequences of actions and decisions
- Implications for organisational practice

Benefits of AI Ethics

Is it only ‘feel good’?

Further benefits, as a subset of corporate ethics,

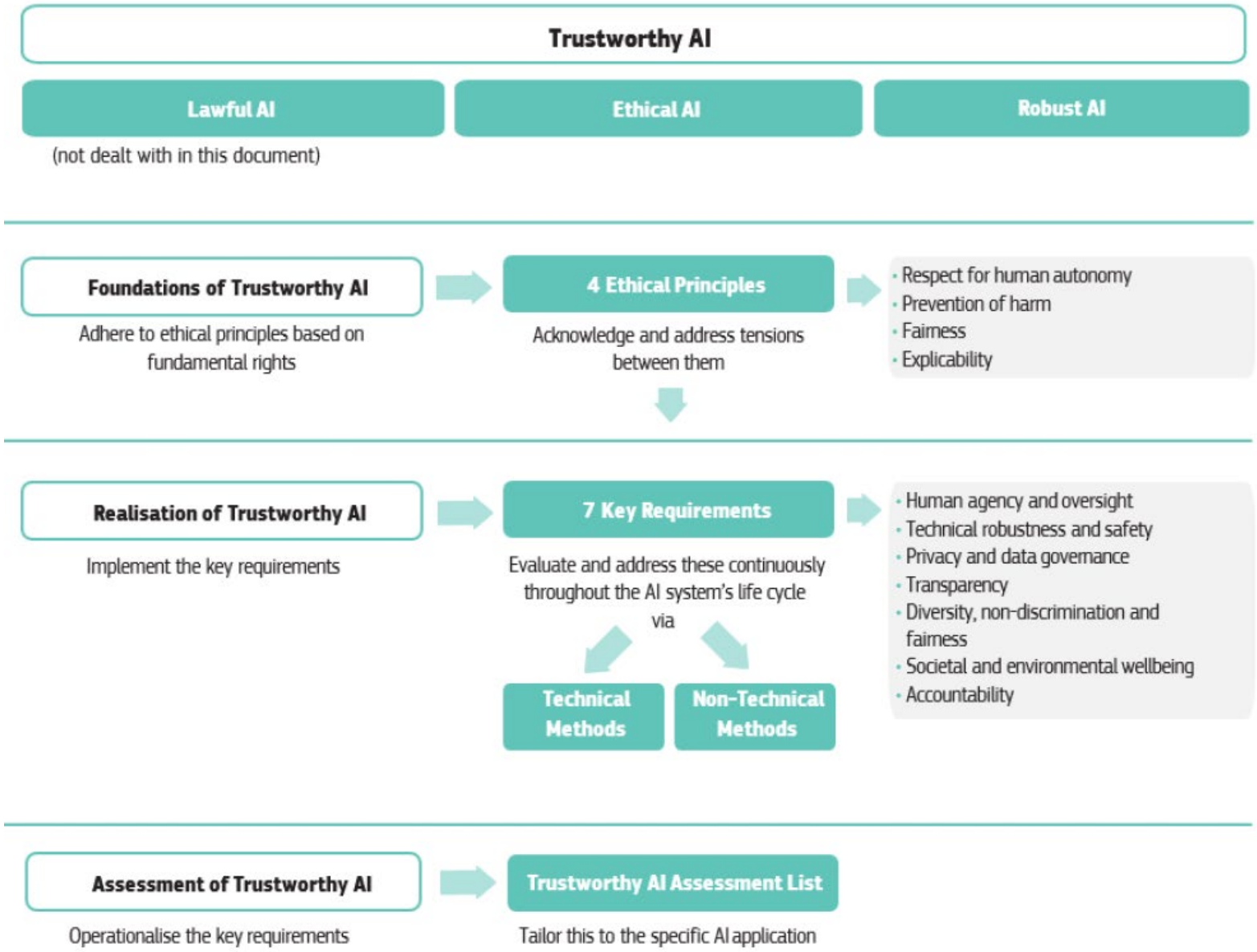
- Positive reputation and brand value
- Employee commitment and morale
- Ease of employee recruiting and retention
- Access to investment capital
- Customer loyalty
- Financial performance?

Trustworthy AI

- EU's High-Level Expert Group on Artificial Intelligence (AI HLEG)
- Mandated with drafting: (1) AI Ethics Guidelines and (2) policy and investment recommendations
- Identified trustworthy AI as the foundational ambition
  - Trust in the design, development, deployment and use of AI
  - The AI technology itself and the socio-technical systems/settings of its application
- Trustworthy AI enables responsible competitiveness
- Three components:
  - Lawful AI – compliance with all applicable laws and regulations
  - Ethical AI – adherence to ethical values, principles, and purpose
  - Robust AI – does not cause unintentional harm

Trustworthy AI - framework

- Foundations – 4 principles
- Realisation – 7 requirements
- Assessment – list for each requirement



Seven requirements

- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, Non-Discrimination and Fairness
- Societal and Environmental Wellbeing
- Accountability

Implementing trust

- Technical and non-technical methods are used to implement the seven requirements:

- Technical
  - Architectures for Trustworthy AI
  - Ethics and rule of law by design
  - Explainable AI
  - Test and validate over multiple metrics
  - Quality of Service indicators

- Non-technical
  - Regulation
  - Codes of conduct
  - Standardisation
  - Certification
  - Accountability via governance frameworks
  - Education and awareness
  - Stakeholder participation and social dialogue
  - Diversity and inclusive design teams

Stakeholders of trust

- All stakeholders have different roles to play in ensuring that the seven requirements are met:
- Owners/ senior management – supervise, report on the requirements to shareholders
- Developers - implement and apply the requirements to design and development processes
- Deployers - ensure that the systems, products and services meet the requirements of trust
- End-users and broader society – know these requirements and able to request that they are upheld

EU AI strategy 2021

- "….with new technologies should not come new values"
- The European Commission published an AI package in April 2021, consisting of:
  - A European Approach for fostering AI
  - A coordinated plan with member states
  - A proposal for AI regulation (EC's White Paper on AI)
- AI regulation has two focus areas (or building blocks): 1) ecosystem of excellence 2) ecosystem of trust
- The ecosystem of excellence concentrates on policy, while the ecosystem of trust focuses on regulation:
  - Motivated by the complexity of AI systems and the risks to human life
  - High risk vs no risk – based on sector level risk and risk in use case
  - Regulations apply case-by-case on seven ethics requirements (next slide)
  - conformity assessments by relevant authorities

EU AI strategy - regulations

- The main focus is on high-risk AI systems – risk is determined through several factors: industry sectors, use and effect (injury, death, or damage). Includes generic AI such as facial recognition
  - Low/no risk AI systems can be subject to this regulation through a voluntary labelling system, which is then binding.
  - Robust and accurate
  - Stakeholder engagement (consumer advocacy, private citizens)
  - Uphold EU values (fairness and balance) in the data used for building the model
  - Documentation, auditing
  - Transparency through explainability
  - Human oversight
- Limitations – public consultation was mostly industrial and tech firms, no worker protections, algorithmic worker surveillance

Ethically Aligned Design (EAD)

- The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
  - Ethically Aligned Design - prioritise human well-being with autonomous and intelligent systems
- Eight principles (read as A/IS will..):
  - Human rights - created and operated to respect, promote, and protect internationally recognized human rights.
  - Well-being - adopt increased human well-being as a primary success criterion for development
  - Data agency - empower individuals with the ability to access and securely share their data, control over their identity.
  - Effectiveness - provide evidence of the effectiveness and fitness for purpose of A/IS.
  - Transparency - The basis of a particular A/IS decision should always be discoverable.
  - Accountability - created and operated to provide an unambiguous rationale for all decisions made.
  - Awareness of misuse - guard against all potential misuses and risks of A/IS in operation.
  - Competence - knowledge and skill required for safe and effective operation.

Other frameworks

- No shortage of guidelines and frameworks
- Regions/countries
  - EU - Ethics Guidelines for Trustworthy AI, 2018
  - USA - Future of Artificial Intelligence, 2016
  - Beijing AI Principles, 2019
  - OECD Principles on AI, 2019
- Organisations:
  - IEEE Ethically Aligned Design, 2019
  - Information Technology Industry AI Policy Principles, 2017
  - Partnership on AI, 2018
  - Google (2018), Microsoft (2019), DeepMind (2019), OpenAI (2018), and IBM (2018)

Limitations of frameworks

- Most recurring - accountability, privacy, fairness, robustness, safety
  - But mostly addressed through technical solutions - FAT ML or XAI community
  - FAMGA - “AI Fairness 360” tool kit, the “What-If Tool”, “Facets”, “fairlearn.py”, “Fairness Flow”
  - Rational only - a typical representation of a male dominated justice ethics (Gilligan 1982)
- Least mentioned -
  - Human autonomy and the superiority of algorithmic decision making
  - Diversity in the discipline and protection of whistle-blowers
  - public–private partnerships – vs industry-funded research (buyout of academic research)
  - Political abuse of AI – “AI race”, automated propaganda, bots, fake news, micro targeting, election fraud
  - Cost of AI – energy usage, carbon emissions, low wage click-workers, lithium mining, technological unemployment

Limitations in practice

- “Ethics plays the role of a bicycle brake on an intercontinental airplane” (Beck 1988)
- Purpose - economic incentive vs societal values or fundamental rights
- Budgets - significant investments in AI for commercial gain vs ethics for public relations purposes
- Staff - neither systematic education of ethical issues, nor empowered to raise ethical concerns
- Approach - voluntary and non-binding, not enforced. An “add-on” to a technical specification.
- Gender - 80% academics and 70% industry are men
- Datasets - collected/curated in the technical domain/developed world on systems/technology built for unrelated objectives

A way forward..

- Micro-ethics - transition from guidelines to instructions
  - Left to the scientist to derive concrete technological implementations from abstract values and principles
  - Datasets - generation, recording, curation, processing, dissemination, sharing and use in training the AI
  - Designing new algorithms with ethics in place, instead of post-processing XAI components
  - Example: standardized datasheets (metadata) listing the properties of training datasets, that can be used to check fit for purpose, original intent, collection, pre-processing.
- Ethical theories – from deontology to virtue ethics
  - Deontologically inspired check-box ticking exercise to a situation-sensitive approach based on virtue ethics
- Increased visibility of ethicists in professional communities, not only the general public.
  - Collaborative public dialogue on the social value of AI

Code of ethics

- Does your 1) company, 2) industry, 3) profession have a code of ethics?
- If there is a code of ethics,
  - Does it address AI/analytics?
  - Does it address or support data and information acquisition, use and organisational conduct?
  - Does it provide a structure to recognize and frame ethical questions?
  - Does it provide a structure to reason through and resolve ethical questions?
  - Does it provide resources to seek ethical guidance?
  - Does it include scenarios to understand ethical positions?

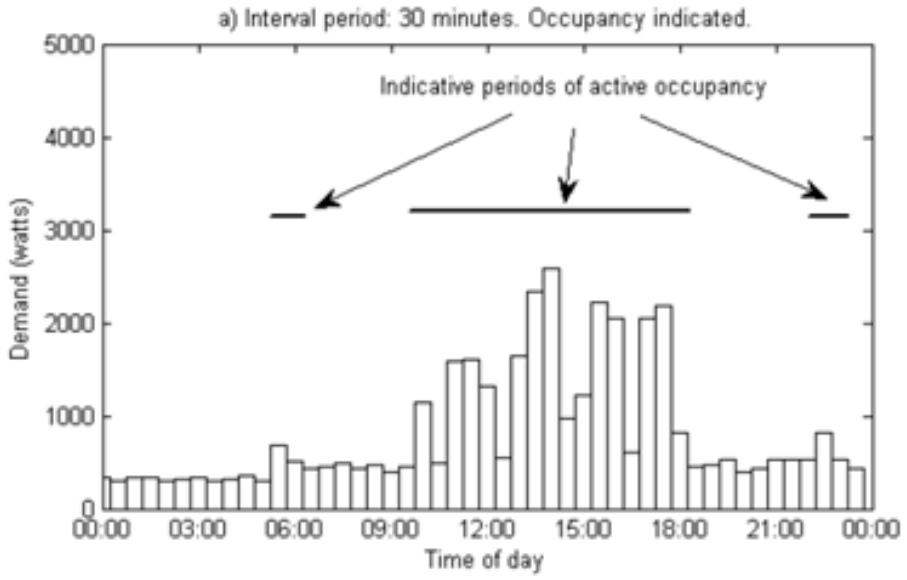
Datasheets for datasets

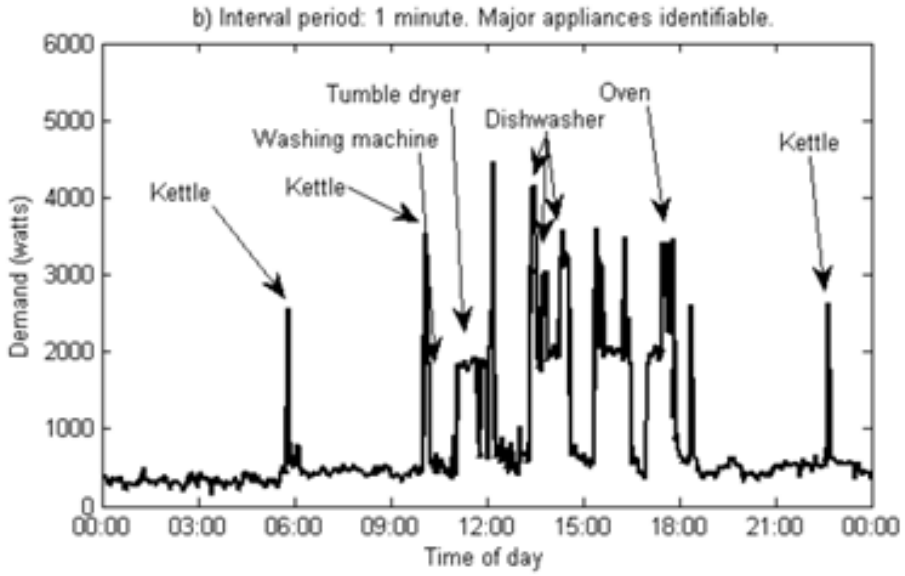
- For dataset creators and dataset consumers
  - Motivation
  - Composition
  - Collection Process
  - Pre-processing/cleaning/labelling
  - Uses
  - Distribution
  - Maintenance

- This article is in LMS ‘Reading’ folder, it contains two worked examples in the appendix.

Example – Smart meters

- Smart meters are used to measure and monitor consumption of utilities (electricity, water, natural gas, and fuel) in near real-time, and transmit a digital data stream of readings to a central node of the grid infrastructure.
- “spy in the home” - which will allow governments to monitor household behaviour
- An abundance of convenience for optimal energy grid management, supplier, distributor, retailer, consumer
- Unlike quarterly or monthly readings of cumulative energy consumption, 15 min or 30 min interval data reveal much more information about the consumers – behaviours and profiles.





Applying Ethics

Application Group	Example Concerns
Illegal uses	Burglars finding out when homes are unoccupied. Stalkers tracking the movements of their victims
Commercial uses	Targeted advertising: Use of individual or aggregated household smart metres data to target advertising at a specific household or individual. Note: Use of aggregated or anonymous data may be more acceptable than use of individual household data. Insurance adjusting e.g. do you tend to leave your appliances on when away from home?
Uses by law enforcement agencies	Detection of illegal activities e.g. sweatshops, unlicenced commercial activities, drug production. Verifying defendant's claim e.g. that they were "at home all evening".
Uses by other parties for legal purposes	In a custody battle: do you leave your child home alone? In a landlord-tenant dispute: is the property over occupied?
Use by family members and other co inhabitants	One householder "spying" on another e.g. parents checking if their children are sleeping or staying up late playing video games. Partners investigating each other's behaviour.

EAD Principle	Implications for Smart Meters/ Smart Grid
Human rights	In a prosumer setting, priorities when charging/discharging from the grid into batteries and EVs.
Prioritizing well-being	In peak shedding events, priority to aged and disadvantaged communities
Accountability	Multi-stakeholder ecosystems from supplier, distributor to retailer
Transparency	Providing consumers full access to their consumption data and comparative information (similar households), across national/local benchmarks
Misuse	Ensuring data and insights are not shared with third parties who may link across other datasets for more individualised profiles

Example – 5G and Edge

- Billions of devices, sensors and IoT that create smart cities, smart transport and smart homes
- Edge computing enables processing/analysing/monitoring/actioning data closer to the origin, 5G enables digital communication with lower latency, higher capacity, and increased bandwidth.
- Network speeds, bandwidth capacities, physical distance, cloud storage are no longer constraints.
- The Edge becomes an intermediary between the source and the cloud, by collecting data, real-time actions, and then de-identifying before being sent to the cloud for further processing and storage.
- This means the original data could be used at the Edge without any impact from data protection laws or privacy concerns.
- Yet, on the cloud, data can be de-anonymised by integrating multiple datasets together and then cross-indexing values against each other
    - COVID related social distancing and contact tracing
    - Pulse of the city using licence plate recognition, crowd movement and traffic monitoring
    - Predictive policing using on-body cameras, squad car cameras, CCTV, social media feeds
    - Edge computing can be used to filter out anomalies like licence plates, accidents or crime

EAD for 5G and Edge

EAD Principle	Implications for 5G and Edge Computing in Smart City settings
Human rights	Full disclosure of surveillance infrastructure currently deployed and how/where it will be used.
Prioritizing well-being	Not only policing or public health measures, but also social equity
Accountability	Multi-stakeholder ecosystems of citizens, organisations and councils
Transparency	Explainable AI and interpretation of insights generated at Edge nodes
Misuse	Well-being measures and not marginalisation of segments of societies

Public Interest Technology

- Public Interest Technology – “a field dedicated to leveraging technology to support public interest organisations and the people they serve” - PIT University Network
- Ensuring data and technology benefit people and society
  - A programmer builds a website that lets parents and their children choose between public schools.
  - A UX expert redesigns a benefits portal so immigrants and counsellors can look at case files together.
  - A data scientist builds a database to encourage the government to start tracking police use of force.
  - Anonymised texting service for teenagers in crisis
- Interdisciplinary, philosophy, ethics, social awareness – scientists, engineers, policymakers, corporate leaders, domain experts, advocacy groups…
- The future of vulnerability: humanity in the digital age
- User-centric design, inclusive use of data and metrics, pilots and prototypes before scaling up