

Ethical and Philosophical Foundations of Digital Humanism

DI Dr. tech. Dr. phil. Erich Prem, MBA
September 2025

1

The good life



Foto: Zac Durant auf Unsplash.

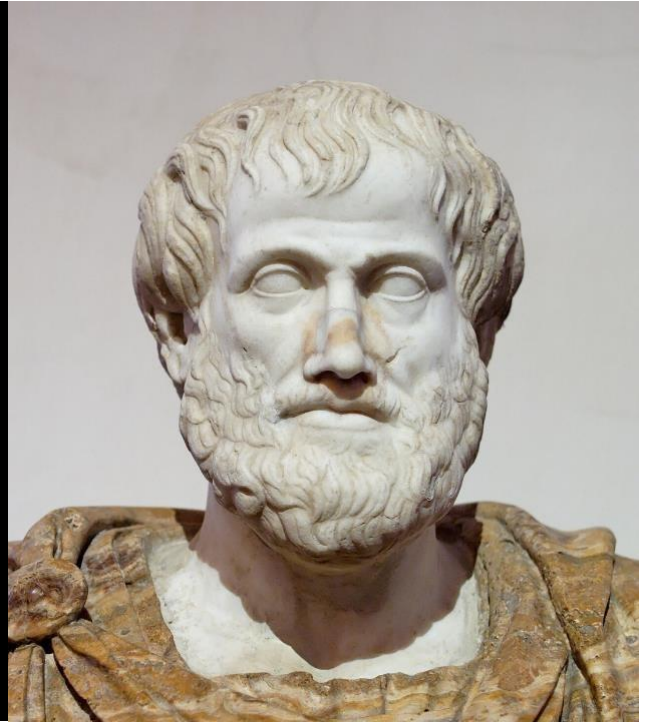
2

Ethics – the good life

“What should I do?”

I. Kant

The goal of all intentional actions is happiness realized in the *good life*.
Aristotle



3

What should I do?

Philosophy of morality

Morality is an informal public system applying to all rational persons, governing behaviour that affects others, and includes what are commonly known as the moral rules, ideals and virtues and has the lessening of evil and harm as its goal.
(Bernard Gert)

εθος – custom (behaviour)

ηθος – character (attitude towards behaviours)

descriptive, normative, applied, metaethics

Some common virtues

truthfulness
courage
honesty
impartiality
reliability
...

Ideals: e.g., justice

Some common harms

death
pain
disability
loss of freedom
loss of pleasure
loss of rights
...

4

Homo mensura

πάντων χρημάτων μέτρον ἐστὶν ἄνθρωπος

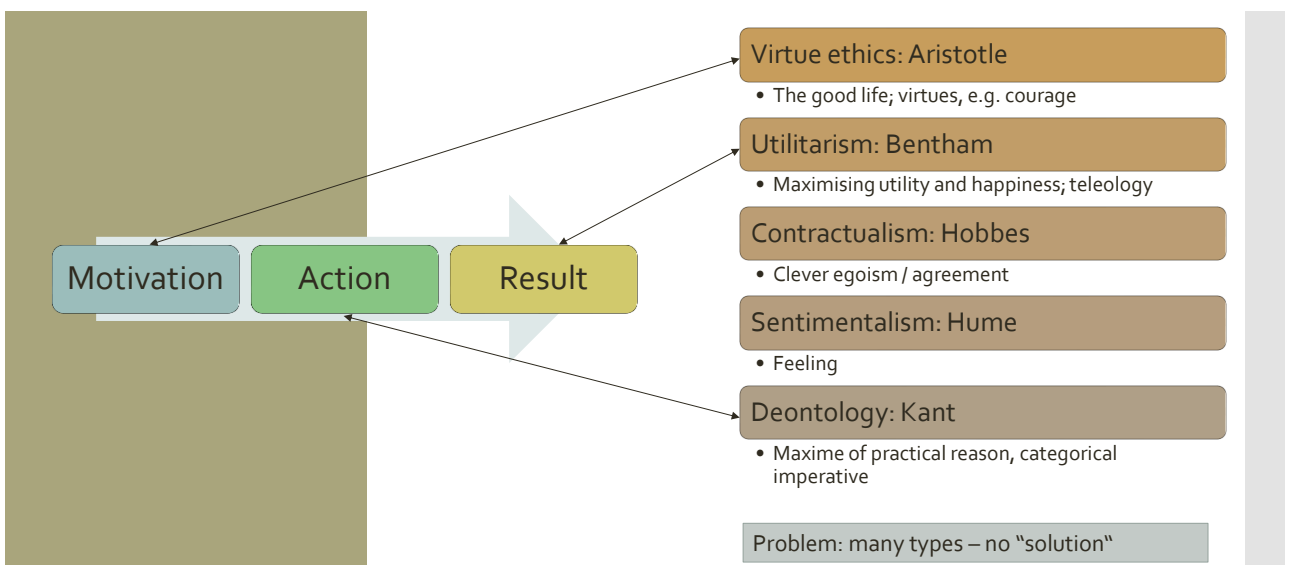
Protagoras, ca. 490-420 v.Chr. (DK 80 B 1)

Digital technology must be designed to empower people and advance our democratic societies.

Digital Humanism, ca. 2019

5

Types of ethics



6

Principlism

The cause of all human evils is not being able to apply general principles to special cases.

Epictetus



7

Large number of “ethics frameworks”...

Table 2 Comparison of ethical principles in recent publications demonstrating the emerging consensus of ‘what’ ethical AI should aspire to be

AI4People (published November 2018) (Floridi et al. 2018)	Five principles key to any ethical framework for AI (L. Floridi and Clement-Jones 2019)	Ethics Guidelines for Trustworthy AI (Published April 2019) (European Commission 2019)	Recommendation of the Council of Artificial Intelligence (Published May 2019) (OECD 2019b)	Beijing AI Principles for R&D (Published May 2019) ('Beijing AI Principles' 2019)
Beneficence	AI must be beneficial to humanity	Respect for human autonomy	Inclusive growth, sustainable development and well-being	Do good: (covers the need for AI to promote human society and the environment)
Non-Maleficence	AI must not infringe on privacy or	Prevention of harm	Robustness, security and safety	Be responsible: (covers the need for researchers to be aware of negative impacts and take steps to mitigate them)
				Control risks: (covers the need for developers to improve the robustness and reliability of systems to ensure data security and AI safety)
				For humanity: (covers the need for AI to serve humanity by conforming to human values including freedom and autonomy)
				Be diverse and inclusive: (covers the need for AI to benefit as many people as possible)
				Be ethical: (covers the need to make the system as fair as possible, minimising discrimination and bias)
				Be ethical: (covers the need for AI to be transparent, explainable and predictable)

Concepts	Basic notions relevant for debating ethical aspects
Principles	Ethical principles (e.g. values)
Concerns	Ways in which principles are threatened through AI systems use and development
Rules	Strategies and guidelines for addressing the challenges

J. Morley et al. (2019) From what to how.
<https://ssrn.com/abstract=3830348>

Explicability AI systems must be understandable and explainable Explicability
 Transparency and explainability Accountability

For a more detailed comparison see Floridi and Cowls (2019) and Hagendorff (2019)



8

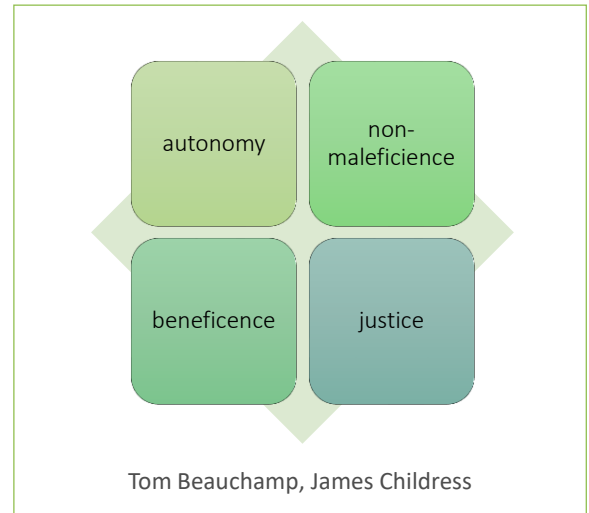
Principlism

Historical core objective:
strengthening personal autonomy

Principle	Example application
Respect for persons	Informed consent
Beneficence	Weighing risks and benefits
Justice	Selection of test subjects

Belmont report (April 18, 1979)

<https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html>



Ethical framework principles

<ul style="list-style-type: none"> Transparency (including explicability, understandability, disclosure etc.) Justice and fairness (including consistency, inclusion, equality, bias, diversity, remedy, redress etc.) Non-maleficence (security, safety, precaution, prevention, integrity etc.) Responsibility (accountability, liability) Privacy 	<ul style="list-style-type: none"> Beneficence (well-being, peace, social good, common good) Freedom & autonomy (consent, choice, self-determination, liberty, empowerment) Trust Sustainability (environment, energy) Dignity Solidarity (social security, cohesion)
--	---

Various possible systems of principles; generally 4-5.

Taken up in politics and regulations, e.g. EU AI Act.

Criticism concerns questions of relevance, governance and how to put them in practice

Agency: a complex concept

Capacity of an actor to act:

For people:

- wilful, intentional action directed at a goal different from reflexes
- question of causation, volition, consciousness etc.

For AI:

- Receive and use data from environment
- Take actions based on input data, autonomously, to achieve goals
- Improve performance by learning from interactions

(Floridi 2023)

AI is less about “intelligence” than about autonomous action (a power to decide)

(Floridi 2023) “a divorce of action and intelligence” because of decoupling successful action from the need to be intelligent and adapting the environment to AI

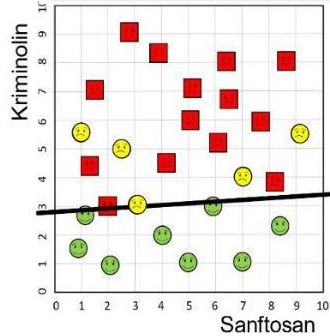
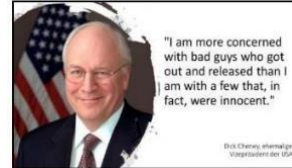
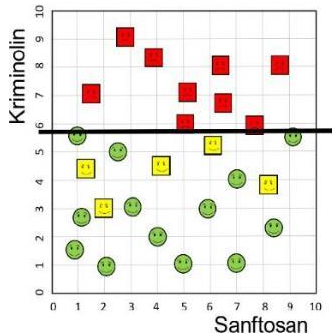
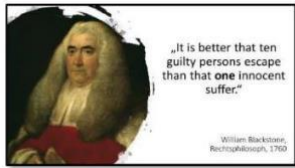
- ➔ Artificial agency
(hence, the question of ethics
and of delegating decisions to automata)

Undesirable bias

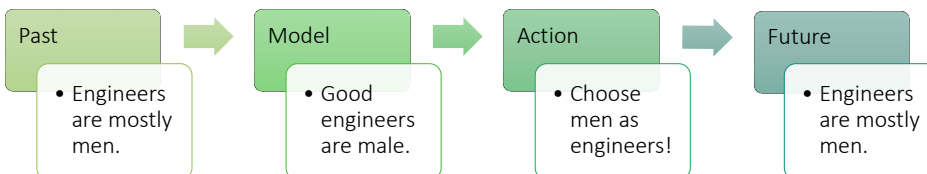
Prejudice is a great time saver. You can form opinions without having to get to the facts.

E.B. White

Discrimination is unavoidable.



Prediction based on the past



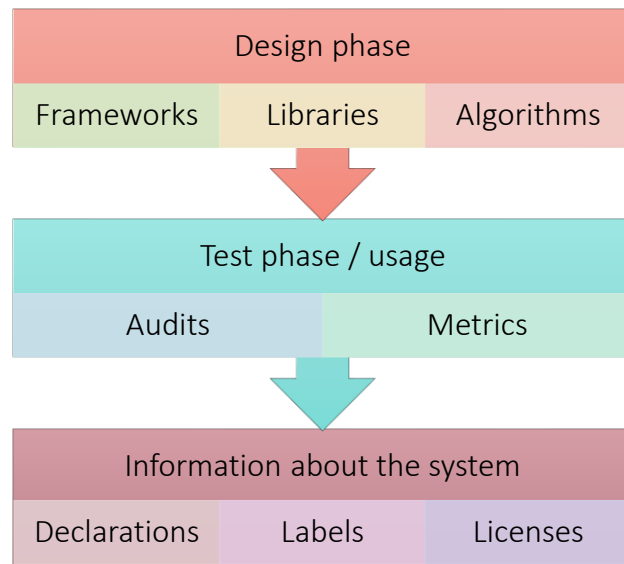
Characterizing different types of unwanted bias

Type of Bias	Description	Examples
Sampling Bias	Occurs when the training data are not representative of the population they serve, leading to poor performance and biased predictions for certain groups.	A facial recognition algorithm trained mostly on white individuals that performs poorly on people of other races.
Algorithmic Bias	Results from the design and implementation of the algorithm may prioritize certain attributes and lead to unfair outcomes.	An algorithm that prioritizes age or gender, leading to unfair outcomes in hiring decisions.
Representation Bias	Happens when a dataset does not accurately represent the population it is meant to model, leading to inaccurate predictions.	A medical dataset that under-represents women, leading to less accurate diagnosis for female patients.
Confirmation Bias	Materializes when an AI system is used to confirm pre-existing biases or beliefs held by its creators or users.	An AI system that predicts job candidates' success based on biases held by the hiring manager.
Measurement Bias	Emerges when data collection or measurement systematically over- or under-represents certain groups.	A survey collecting more responses from urban residents, leading to an under-representation of rural opinions.

Approaches, tools, methods

...and many open issues.

Tools and methods for various design phases



From what to how: recurring issues

Summaries	Notions	Procedures	Code	Infrastructure	Education	Ex-post assessment and agreement
Overviews and introductions	Frameworks and concepts	Process models	Algorithmic methods	Data sets	Training and tutorial	Audit
Case studies and examples	Criteria and checklists	Guidelines and codes of practice	Design patterns	Online communities		License model
	Declarations	Standards	Software libraries			
	Metrics		Software assistants			
Good practice	Regulation	Consulting		Ethics councils and boards	Coaching	Labels, warnings, consent management

Labels provide information about AI models

Inspiration from labels for food,
clothing for consumers:
“model cards”

Shift of responsibility to the users.

Fiction of consent: experience from

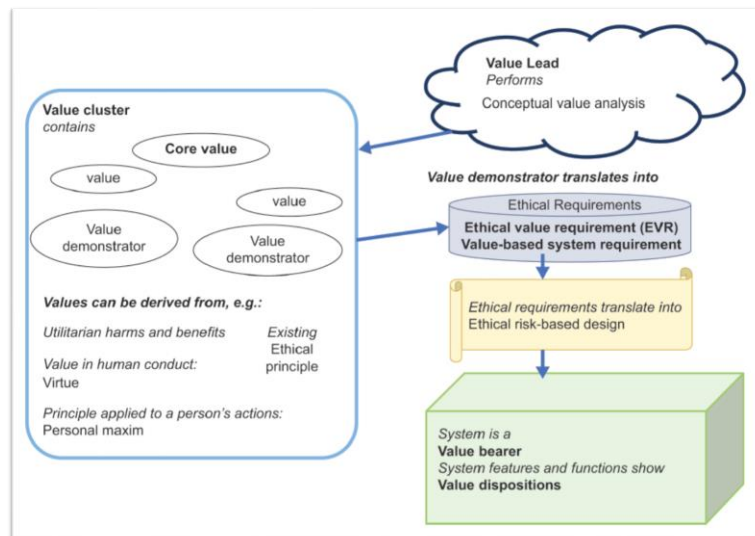
- Terms of Use
- Dark Patterns / GDPR agreement
- etc.



Standards

Existing standards for
AI/autonomous systems

- Model process for addressing ethical concerns during systems design (IEEE 7000-2021)
- Transparency of autonomous systems (IEEE 7001-2021)
- Data privacy process (IEEE 7002-2022)
- Algorithmic bias considerations (IEEE P7003)
- Standards on child and student data governance (IEEE P7004)
- ...



AI bias mitigation & associated challenges

Ferrara, E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* **2024**, 6, 3. <https://doi.org/10.3390/sci6010003>

Approach	Description	Examples	Limitations and Challenges	Ethical Considerations
Pre-processing Data	Involves identifying and addressing biases in the data before training the model. Techniques such as oversampling, undersampling, or synthetic data generation are used to ensure the data are representative of the entire population, including historically marginalized groups.	1. Oversampling darker-skinned individuals in a facial recognition dataset [1]. 2. Data augmentation to increase representation in underrepresented groups. 3. Adversarial debiasing to train the model to be resilient to specific types of bias [33].	1. Time-consuming process. 2. May not always be effective, especially if the data used to train models are already biased.	1. Potential for over- or underrepresentation of certain groups in the data, which can perpetuate existing biases or create new ones. 2. Privacy concerns related to data collection and usage, particularly for historically marginalized groups.
Model Selection	Focuses on using model selection methods that prioritize fairness. Researchers have proposed methods based on group fairness or individual fairness. Techniques include regularization, which penalizes models for making discriminatory predictions, and ensemble methods, which combine multiple models to reduce bias.	1. Selecting classifiers that achieve demographic parity [31]. 2. Using model selection methods based on group fairness [11] or individual fairness [30]. 3. Regularization to penalize discriminatory predictions. 4. Ensemble methods to combine multiple models and reduce bias [34].	Limited by the possible lack of consensus on what constitutes fairness.	1. Balancing fairness with other performance metrics, such as accuracy or efficiency. 2. Potential for models to reinforce existing stereotypes or biases if fairness criteria are not carefully considered.
Post-processing Decisions	Involves adjusting the output of AI models to remove bias and ensure fairness. Researchers have proposed methods that adjust the decisions made by a model to achieve equalized odds, ensuring that false positives and false negatives are equally distributed across different demographic groups.	Post-processing methods that achieve equalized odds [11].	Can be complex and require large amounts of additional data [32].	1. Trade-offs between different forms of bias when adjusting predictions for fairness. 2. Unintended consequences on the distribution of outcomes for different groups.

21

Fair training methods

Fair Training Method	Definition	Implementation	Key Features	References
Pre-processing Fairness	Modifying training data before feeding into the model	Re-sampling, re-weighting, data augmentation	Addresses bias at the data level	[136,139,140]
In-processing Fairness	Modifying learning algorithms or objective functions	Adversarial training, adversarial debiasing	Simultaneously optimizes for accuracy and fairness	[137,141,142]
Post-processing Fairness	Adjusting the model's predictions after training	Re-ranking, calibration	Does not require access to the model's internals	[46,143–145]
Regularization-based Fairness	Adding fairness constraints to the optimization process	Penalty terms in the loss function	Can be combined with various learning algorithms	[43,146,147]
Counterfactual Fairness	Measuring fairness based on changes in sensitive attributes	Counterfactual reasoning	Focuses on individual-level fairness	[45,148,149]

22

The trouble with fairness

...and other principles.



23

Where does fairness arise in engineering?



- Modelling
- Categorization
- Classification
- Prediction

Algorithmic
decision-making
including AI



- Autonomous driving
- Robots
- Adaptive interfaces

Autonomous
action – AI
Agency



From doorknobs to user interfaces, fairness arises everywhere!



<https://www.melaninbasecamp.com/trip-reports/2024/5/17/what-is-hostile-architecture-americas-war-on-the-unhoused>

24

24

Which discrimination...is fair?



Ethical discrimination

- Certain characteristics should not result in disadvantages (often they have in the past)
 - ethnicity, gender, religion, age, disability, sexual orientation
 - often targets a change in society (policies)
- Distinction between inequalities for explainable discrimination
 - Income: relevant feature
 - Gender: irrelevant

In practice very difficult!
- Modern proposal: include only attributes that an individual can *directly influence*, e.g. no one should be treated worse just out of bad luck.

Inequality type	Example
Natural	Disability at birth
Socioeconomic	Parents' assets
Talent	Skills
Preference	Saving behaviour
Treatment	Job market discrimination

For example, biases: what is *really* fair?

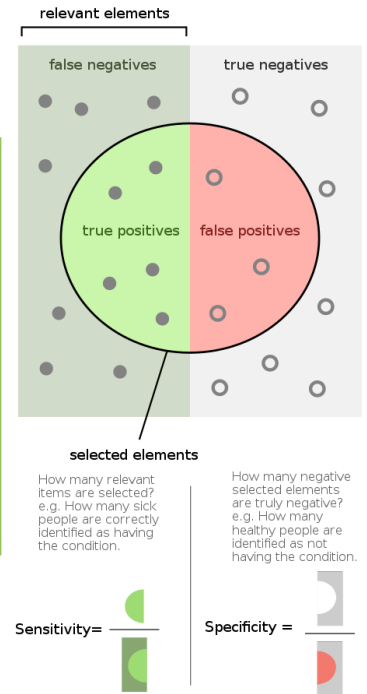
Assume: modelling default risk of a lender on a loan.

Scenario: supervised learning, some “inappropriate” attribute present, e.g. race, gender, social status

- False positives (FP): lost opportunity (predicted default, but would have repaid)
- False negative (FN): lost revenue (predicted repayment, but defaulted)

Various error rates:

- True positive rate, sensitivity, probability that an actual positive will test positive: $(TPR)=TP/(TP+FN)$
- True negative rate, specificity: $(TNR)=TN/(FP+TN)$
- False positive rate, fall-out: $(FPR)=FP/(FP+TN)=1-TNR$
- False negative rate (FNR)= $FN/(FN+TP)=1-TPR$
- Positive predictive value, precision: $(PPV)=TP/(TP+FP)$



M. Seng Ah Lee, L. Floridi, J. Singh (2021) Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. <https://ssrn.com/abstract=3679975>

27

Which inequality is fair? A selection of ideas...

Fairness metric (literature)	Equalising	Intuition/example	
Maximise total accuracy	N/A	Most accurate model gives people the loan and interest they 'deserve' by minimising errors	desert*
Demographic parity, group fairness	Outcome	Black and white applicants have same loan approval rates	strict egalitarianism
Equal opportunity	FNR	Among creditworthy applications, black and white applicants have similar approval rates	Fair equality of opportunity
Predictive equality	FPR	Among defaulting applicants, black and white have similar rates of denied loans	
Equal odds	TPR, TNR, PPV	Both of the above: Among creditworthy applicants, probability of predicting repayment is the same regardless of race	
Counterfactual fairness	Prediction in counterfactual scenario	For each individual, if they were a different race, the prediction would be the same	Cause and effect
Individual fairness	Outcome for 'similar' individuals	Each individual has the same outcome as another 'similar' individual of a different race	Responsibility-sensitive egalitarianism

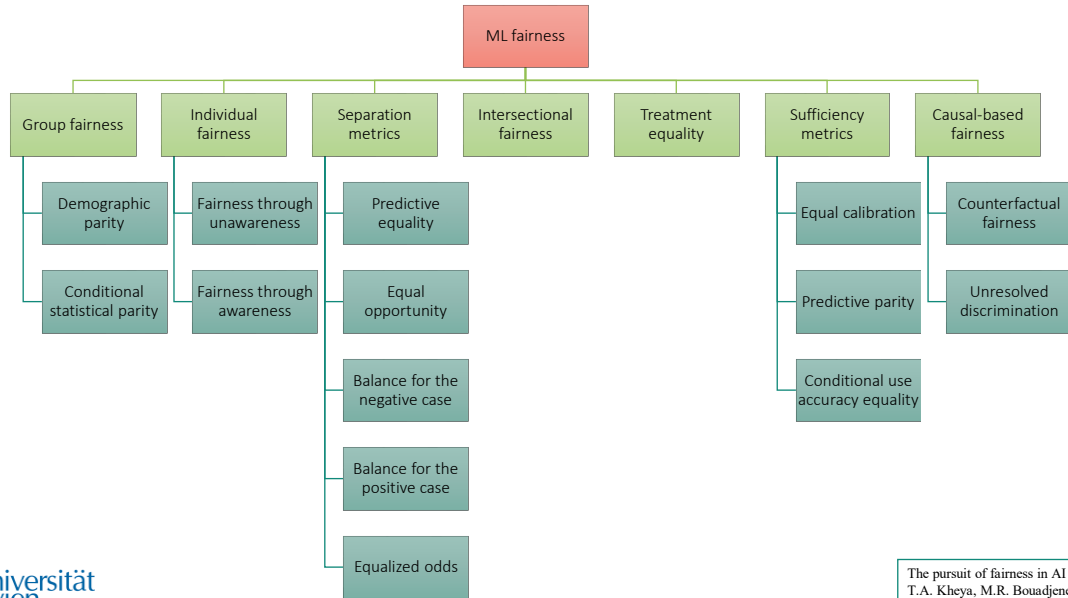


M. Seng Ah Lee, L. Floridi, J. Singh (2021) Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. <https://ssrn.com/abstract=3679975>

* What one deserves, not the dry landscape.

28

Fairness concepts



Types of fairness definitions

Different basic ideas about fairness – typically mixed in human practice



Type of Fairness	Description
Group Fairness	Ensures that different groups are treated equally or proportionally in AI systems. Can be further subdivided into demographic parity, disparate mistreatment, or equal opportunity.
Individual Fairness	Ensures that similar individuals are treated similarly by AI systems, regardless of their group membership. Can be achieved through methods such as similarity-based or distance-based measures.
Counterfactual Fairness	Aims to ensure that AI systems are fair, even in hypothetical scenarios. Specifically, counterfactual fairness aims to ensure that an AI system would have made the same decision for an individual, regardless of their group membership, even if their attributes had been different.
Procedural Fairness	Involves ensuring that the process used to make decisions is fair and transparent.
Causal Fairness	Involves ensuring that the system does not perpetuate historical biases and inequalities.

Inherently political

Limits qua design	Limits of mathematical fairness	Bias may be persistent.	De-biasing can be costly.	Bias paradox
<ul style="list-style-type: none"> • All technical systems are value-laden. 	<ul style="list-style-type: none"> - Some reasonable expectations about fairness may not be realizable in parallel (sufficiency-separation debate). 	<ul style="list-style-type: none"> • We may decide to use biased systems, e.g. in medicine. 	<ul style="list-style-type: none"> • Increasing the likelihood of granting a loan for women may cause more loan defaults for all and higher rates for everybody (at least if women default more than men). 	<ul style="list-style-type: none"> - No forbidden attributes for decision-making. - But change in society (policies) based on "forbidden" attributes.

Context matters

- Affirmative action to grant women easier access to university
- Past disadvantage, skewed data

University admission



- Past disadvantage
- No tolerance regarding training and tests
- Different affirmative action needed

Pilot license



Limits of algorithmic decidability

Amor vincit omnia: art, medicine or porn?

- Is pornography reducible to nudity?

Externalisation / extensionalisation

- Digital means to assess legality
- Extension ethics:
exclusive orientation at external (formal) criteria.
- Assessment based on appearance, not on intent. Intentions are never depicted.
- Issue of context and form reducibility



https://de.wikipedia.org/wiki/Datei:Caravaggio_-_Cupid_as_Victor_-_Google_Art_Project.jpg

33

33

AI risks

Doing business with AI



34

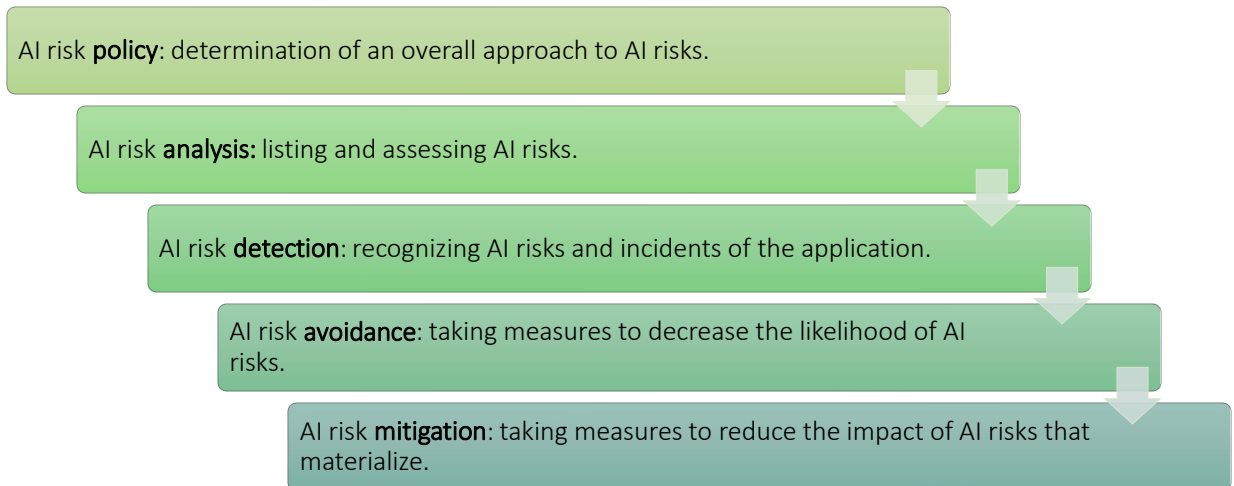
AI Business Risks

Type of risk	Examples
Innovation risks	Non-acceptance of new technologies, missed opportunities through technology lock-in
Technology risks	Lack of robust and correct functioning, threats from new and better technologies, quality issues, residual risks due to lack of proven correctness
Security risks	Cybersecurity attacks, data leaks, unpredicted and dangerous behaviour
Public relations risks	Lack of credibility towards consumers, trustworthiness of products and services, public perception of unfair treatment and bias
Regulatory risks	Changes in regulatory environment that impact on value proposition, production costs, work environments etc.
Legal risks	Challenging new regulatory requirements, lawsuits emerging from legal conflicts
Human resources risks	Deskilling, lack of staff at required skilled levels
Market risk	Unfulfilled user expectation due to hyped AI technology perception

Business risks emerging from AI ethics issues

Ethical issue (principle)	Reaction from public, partners & consumers	Business risk	Risk avoidance and mitigation strategies
Lack of transparency (transparency)	Distrust, resistance	Sales loss, missed opportunities, legal and regulatory issues	Open data and process, model card (information), processes for providing explanations, human intervention, and oversight
Bias, discrimination (fairness)	Distrust, complaints	Negative public perception, complaint management	Debiasing, diversity measures, testing, user information, industry standards (e.g. fairness)
Privacy infringement (privacy)	Consumer complaints, resistance, distrust	Complaint management, lawsuits (e.g. GDPR), sales loss	Safe data handling practices, improved privacy technologies, minimization of data needs, preparation for data losses
Security risks (non-maleficence)	Compensation requests, distrust	Legal procedures and lawsuits, negative public perception, complaint management	Quality assurance, testing, monitoring, early detection, maintenance
Regulatory non-compliance	Distrust, complaints, public inquiry	Negative public perception, legal costs	Compliance processes, monitoring, audits, early detection, maintenance
Misinformation, manipulation, system abuse	Public complaints & calls for action, political attention	Negative public perception, PR costs, change of technology or business model	Monitoring, early detection, legal procedure, public statements, contract management
Concentration of power (own)	Distrust, monopoly action	Legal procedures, limitation in choice of partners, premium prices, service restrictions	Establish relationship and communication with regulator

Managing AI Risks



Risk mitigation strategies

Early detection: Often the early detection of incidents can help address problems and save costs as well as further consequences or more incidents. Relevant tools include monitoring, feedback channels for users, documentation, and reporting.

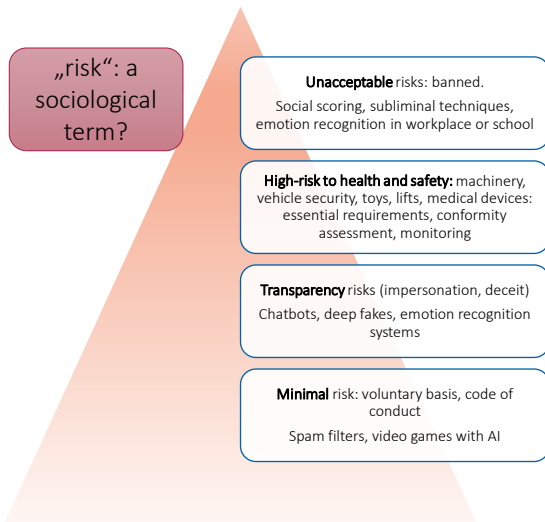
Minimising financial consequences: This may include preparations for rapid discontinuation of services, insurance for business and customers, switching to different or previous systems.

Ensuring consumer trust: Provision of updates and repairs, procedures for maintenance, information of consumers such as handbooks and self-help guidelines.

Minimising negative public perception: Companies should be prepared to address incidents publicly. This can include an explanation what happened, communication how the situation was addressed and how it will be avoided in the future. It can include an investigation and public statements from top management executives.

Policy links: AI businesses should establish functioning links with policy makers and industry associations, in particular in areas of identified risks. These links can be useful to steer regulation, to establish commonly accepted industry standards and to address incidents at a more general level beyond that of just a single company.

The famous risk-based approach

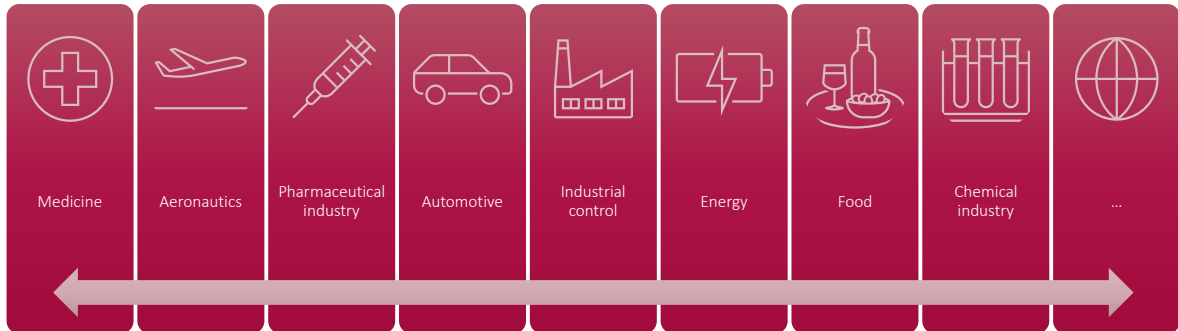


- Complex legal environment, e.g. GDPR, product safety, IPR, competition law, sector regulation, consumer rights etc.
- Compliance rules for companies
 - Risk management, data governance (e.g. training data), record keeping, transparency (to users to interpret the systems), human oversight, accuracy, robustness, cybersecurity
- A lot of critique from legal scholars
 - Definitions, complex differentiation (e.g. provider intentions), formulations, reliance on internal reporting duties

Examples

Forbidden	High-risk	Transparency risks	GPAI
Chapter II, Art. 5 <ul style="list-style-type: none"> • Real-time remote biometric identification in public spaces • Social scoring systems • Subliminal influencing exploiting vulnerabilities of specific groups • Assessing criminal offence risks • Inferring emotions in the workplace • ... 	Chapter III <ul style="list-style-type: none"> • Safety components • Biometrics (except just verification) • Critical infrastructure • Education / training • Employment • Essential public services • Law enforcement • Borders and migration • Justice • Profiling of individuals for assessment (e.g. work) 	<ul style="list-style-type: none"> • Chat bots • Deep fakes 	General-purpose AI <ul style="list-style-type: none"> • E.g. Chatbots, foundational models • Documentation requirements • Respecting the Copyright Directive • Special rules for free and open GPAI models •

Learning from risky business



Pharmaceutical industry

Process-level controls: clear and formal rules with strict controls for the production processes

Auditing: high levels of internal and external auditing: consultants, regulators. Audits are frequent and for different processes. Internal control by teams that receive special training. Severe threats to businesses not following up on observations.

Quality standards: very clear rules about acceptable quality deviations and documentation issues and follow-up procedures. International standards with effective follow-up. A whole set of actors with reporting and follow-up duties.

Specialised staff: Quality management by trained staff with knowledge in pharmaceuticals, production & engineering, and the regulatory environment on top of an understanding of business issues.

© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com



"I didn't experience any of the side effects listed in the enclosed literature. Should I be concerned?"

search ID: mbcn2001

Aeronautics safety

Monitoring and maintenance: clear and concise but useful routines for checking and maintaining operational safety. Well-established practices from short checks before every flight to routine complete overhauls of planes.

Error handling: systems to reporting incidents and potential safety issues, continuous improvement of safety evolved over decades. Systematic investigations after incidents leading to information, warnings, recommendations, or grounding.

Redundancy: includes not fully relying on the actions of others and to perform systematic double-checking as well as safe fall-back procedures in case of system failures.

Safety culture: a principle of "safety first" to reporting and whistleblowing, programs that facilitate the confidential reporting of errors, shortcomings, and malfunctions, their potential causes and how to address them



Ladies and gentlemen, this is your captain speaking.
There is a minor malfunction in the pressurization system, but no problem, an oxygen mask will come out of the unit above your seat automatically

Situation prototypes

Being good like before



Medical ethics prototype situations



Ethical
principles

Regulation

Code of
practice

Teaching

Shared
objectives

Tradition

Prototype situations

School admission	Credit rating	Recruitment
Music recommendation	Product recommendation	Partner recommendation
News recommendation	Dynamic pricing	Fraud detection
Fitness assistant	Language learning	...



Law, principles, frameworks

Generalization, Specialisation from
tradition and non-digital approaches

Virtue in the digital realm

Being good



Jakub Geltner Nest 05 (2015)



47

Dr.phil. Dr.tech. Erich Prem (MBA)

www.erichprem.at

prem at eutema.com

@ErichPrem

Institut für Philosophie, Uni Wien

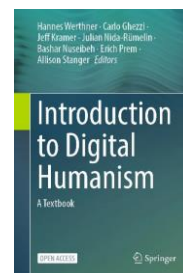
<https://philtech.univie.ac.at/team/erich-prem/>

eutema GmbH

www.eutema.com

www.digitalhumanism.at

<https://link.springer.com/book/10.1007/978-3-031-45304-5>



<https://dighum.ec.tuwien.ac.at/perspectives-on-digital-humanism/>

Contact



48

48