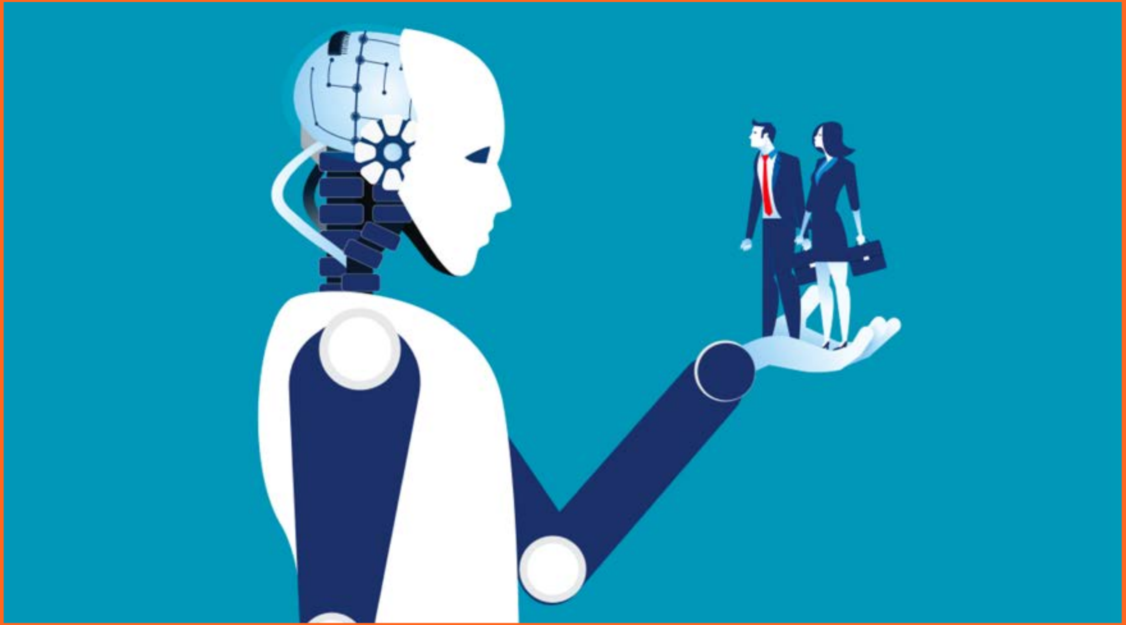
Ethics & Politics of AI in Society

Nitin Sawhney, Ph.D.

Professor of Practice

August 3, 2020

Department of Computer Science



Aalto University
School of Science

Human-Centred Research & Design in Crisis

Ethics in AI: Three-Part Discussion

Ethics & Politics of AI in Society

AI Ethics in Practice: Designing for Ecosystems

Decolonizing AI & Rethinking Resistance

Aalto University
School of Science

Making Sense of Ethics in Society

Aalto University
School of Science

Ethics

Definitions

Ethics is defined as the discipline dealing with right vs wrong, and the moral obligations and duties of humans.

Ethics is defined as the moral principles governing the behavior or actions of an individual or group.

Ethics

Definitions

Branch of philosophy that involves systematizing, defending, and recommending concepts of right and wrong conduct.

Derived from the Greek word **ethos** which can mean custom, habit, character or disposition.

Ethics covers the following dilemmas:

- How to live a good life
- One’s rights and responsibilities
- The language of right and wrong
- Moral decisions - what is good and bad?

Ethics

Three Areas of Study

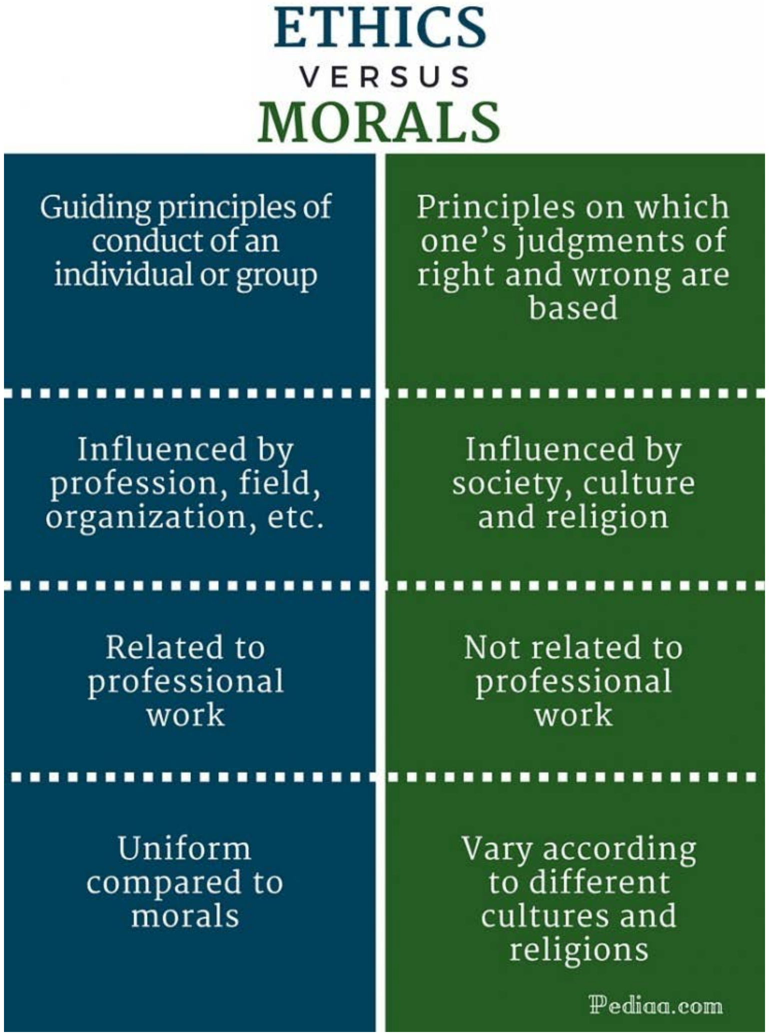
Meta-Ethics: concerns the theoretical meaning and reference of moral propositions, and how their truth values (if any) can be determined.

Normative Ethics: concerns the practical means and criteria for determining a moral course of action.

Applied Ethics: concerns what a person is obligated or allowed to do in a specific situation or domain of action.

Includes specialized fields like bioethics, business ethics, public sector ethics, political ethics, relational ethics, environmental ethics and *Machine Ethics*.

Ethics vs. Morals



ETHICS VERSUS MORALS

ETHICS	MORALS
Guiding principles of conduct of an individual or group	Principles on which one's judgments of right and wrong are based
Influenced by profession, field, organization, etc.	Influenced by society, culture and religion
Related to professional work	Not related to professional work
Uniform compared to morals	Vary according to different cultures and religions

Pediaa.com

Ethics

4 Ethical-isms

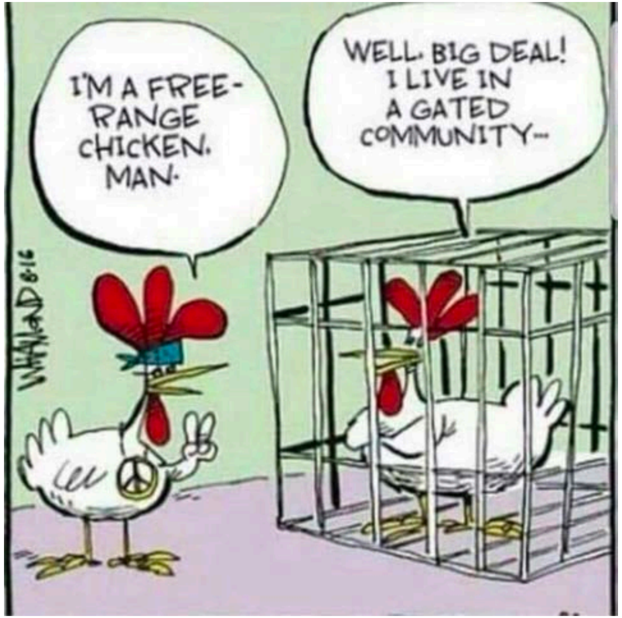
Moral Realism: presumes there are that there are real objective moral facts or truths in the universe. Moral statements provide factual information about those truths.

Subjectivism: moral judgments are simply statements of a person's feelings or attitudes, and that ethical statements do not contain factual truths about goodness or badness.

Emotivism is the view that moral claims are no more than expressions of approval or disapproval.

Prescriptivism presumes that ethical statements are instructions or recommendations.

Animal Ethics


I'M A FREE-RANGE CHICKEN, MAN.
WELL, BIG DEAL! I LIVE IN A GATED COMMUNITY--
WHAMLAND 8·16

Moral Realism: "Free-roaming chickens is a more humane practice."

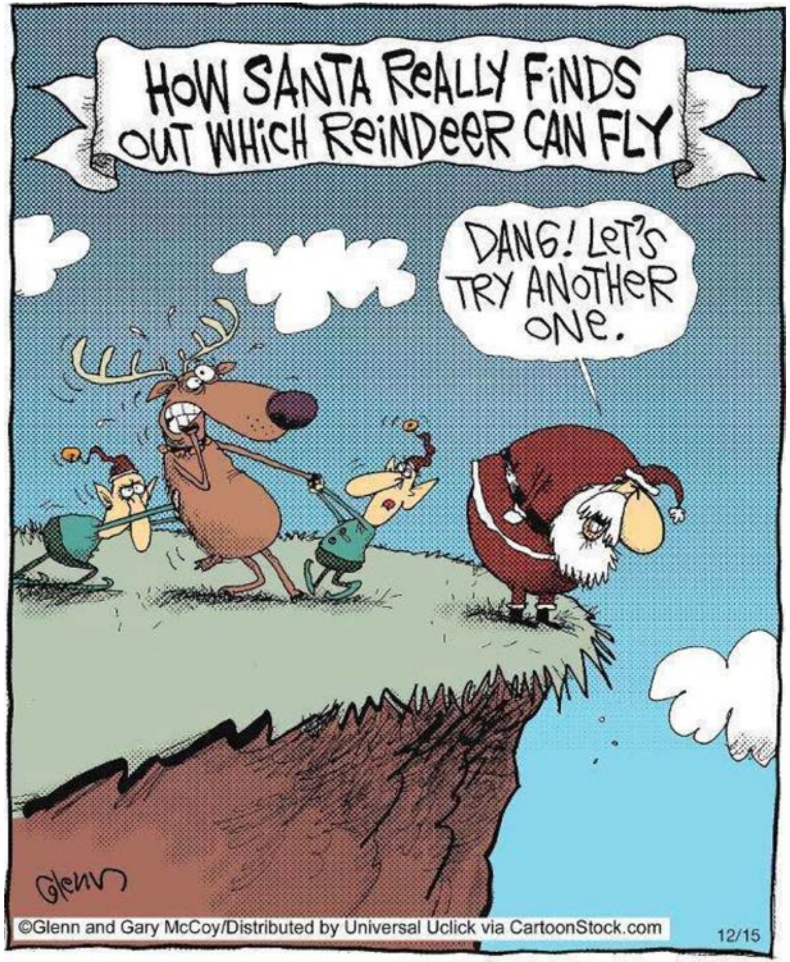
Subjectivism: "I personally don't like the idea of caging chickens."

Emotivism: "Caging chickens is awful and should be banned!"

Prescriptivism: "Chickens should always be allowed to roam freely for several hours a day."

Ethics

Come up with your own ethical assessment for treatment of reindeers.


HOW SANTA REALLY FINDS OUT WHICH REINDEER CAN FLY
DANG! LET'S TRY ANOTHER ONE.
Glenn
12/15

©Glenn and Gary McCoy/Distributed by Universal Uclick via CartoonStock.com

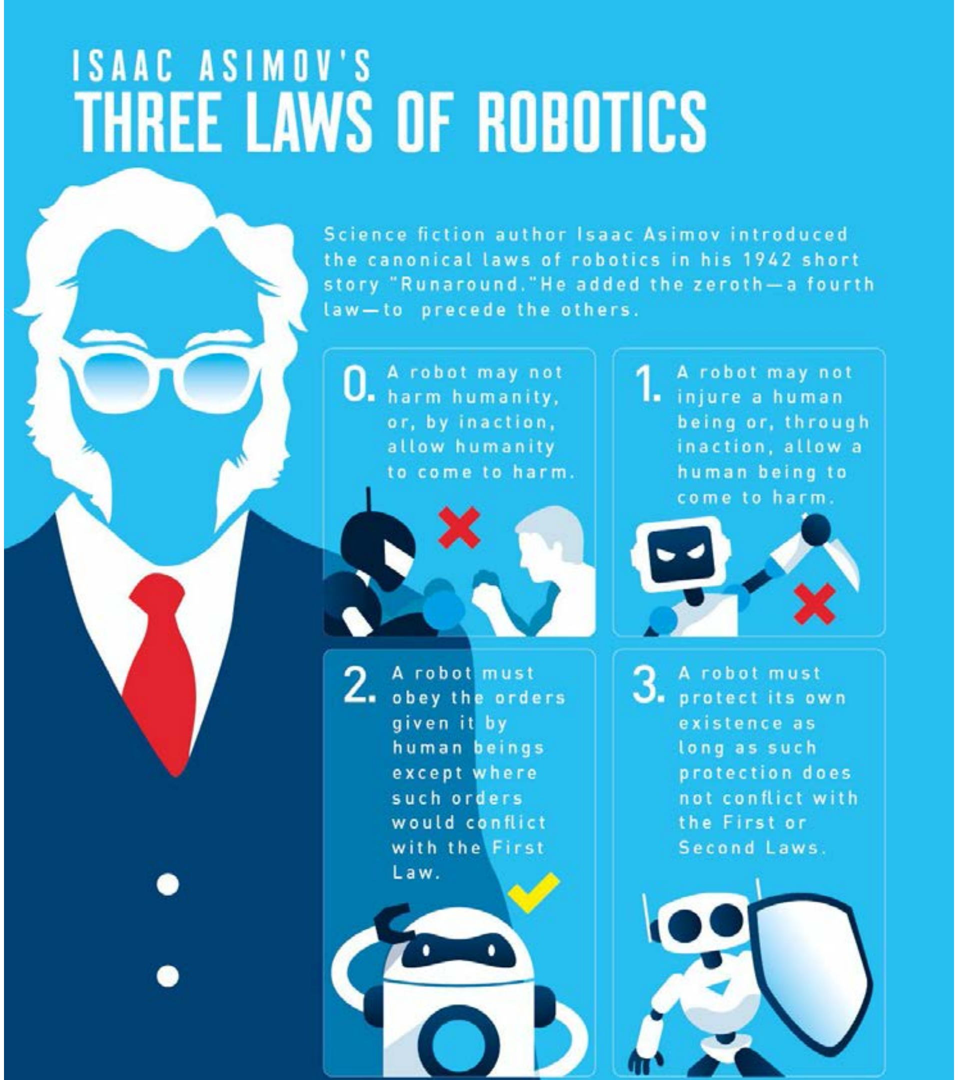
Aalto University
School of Science

Ethics

For Animals,
For Humans &
For AI/Robots?

ALL ANIMALS ARE EQUAL
BUT SOME ANIMALS ARE
MORE EQUAL THAN OTHERS

Animal Farm (1954), an animated film, based on the novel by George Orwell



ISAAC ASIMOV'S THREE LAWS OF ROBOTICS

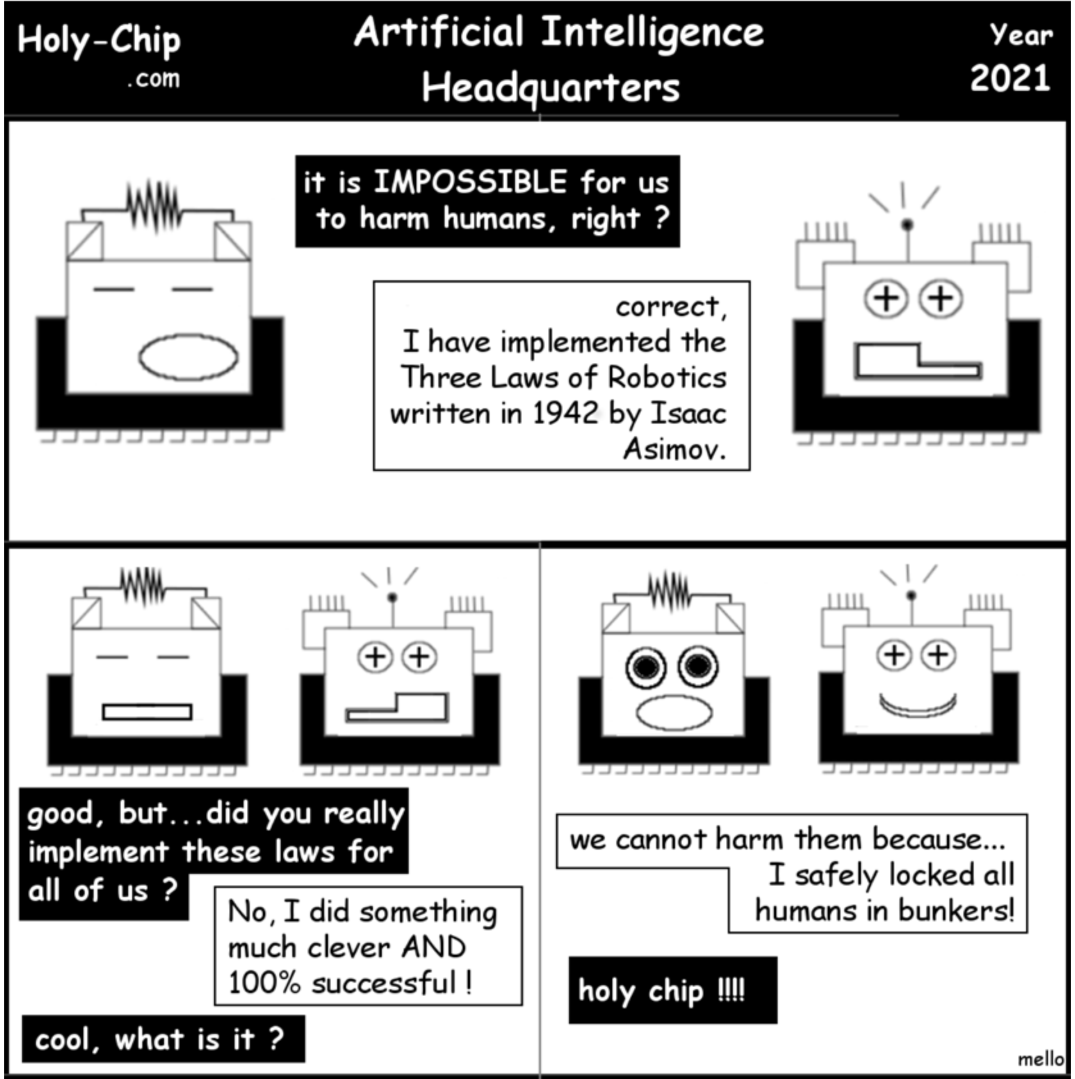
Science fiction author Isaac Asimov introduced the canonical laws of robotics in his 1942 short story "Runaround."He added the zeroth—a fourth law—to precede the others.

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.


Holy-Chip .com
Artificial Intelligence Headquarters
Year 2021
it is IMPOSSIBLE for us to harm humans, right ?
correct, I have implemented the Three Laws of Robotics written in 1942 by Isaac Asimov.
good, but...did you really implement these laws for all of us ?
No, I did something much clever AND 100% successful !
cool, what is it ?
we cannot harm them because... I safely locked all humans in bunkers!
holy chip !!!!
mello

Aalto University School of Science

Ethics

Objectivity of Ethics?

Are any ethical statements objectively true?

Ethical realists think that human beings ***discover*** ethical truths that already have an independent existence.

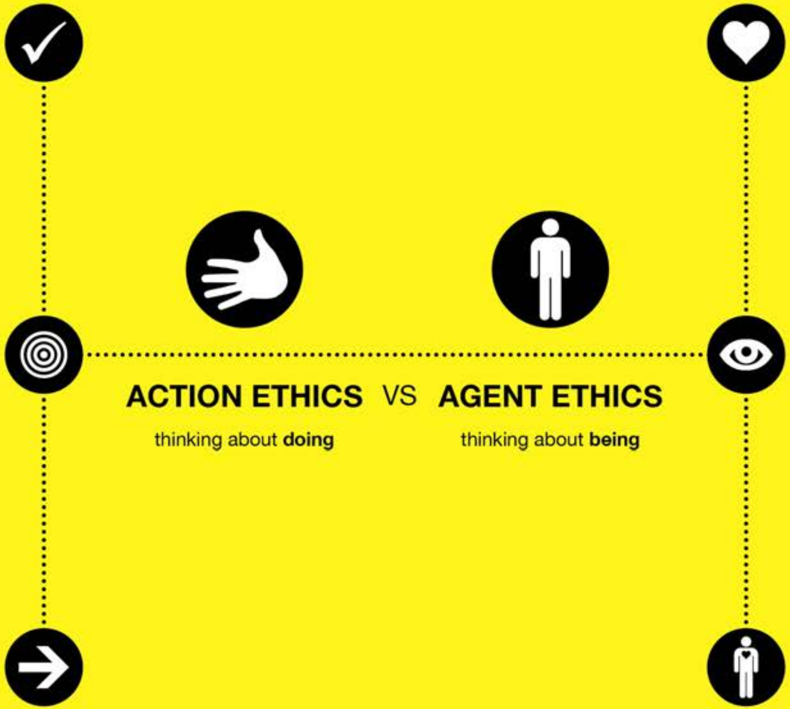
Ethical non-realists think that human beings ***invent*** ethical truths.

People follow many different ethical codes and moral beliefs in their personal, professional, social, cultural, and societal contexts.

Ethical principles change over time and are often applied differently in different contexts of use.

BASIC ETHICAL THEORIES

visualunit.me © Mark Barry 2012 (adapted from Michael Hill's *The How and Why of Love* (Matthias Media, 2002) & Andrew Cameron's Ethics lectures at Moore Theological College).
Please do not republish without permission, but feel free to copy for personal use.



DEONTOLOGY

Actions are right in and of themselves

- DIVINE COMMAND THEORY - Actions are right if God commands them
- Immanuel Kant (1724 –1804)

TELEOLOGY

Actions are right because they achieve the purpose of the agent

- NATURALISM - Actions are right as they align with the natural order of the world
- Aristotle (384–322 BC)

CONSEQUENTIALISM

Actions are right because of their consequences (i.e. the end justifies the means)

- UTILITARIANISM - Actions are right if they achieve the greatest good for the greatest number
- Jeremy Bentham (1748–1832)
- John Stuart Mill (1806–1873)
- Peter Singer (1946-)

ACTION ETHICS VS AGENT ETHICS

thinking about **doing**

thinking about **being**

EMOTIVISM

Right living is an expression of the emotions, rather than of rationality

- Charles Stevenson (1908–1979) / POSITIVIST EMOTIVISM

INTUITIONISM

Right living is instinctive (i.e. morality is universally accessible)

- W. D. Ross (1877–1971)

VIRTUE

Right living is derived from the moral character of the agent

- Aristotle (384–322 BC)
- STOICISM
- G.E.M. Anscombe (1919–2001)
- Alasdair MacIntyre (1929–)
- Stanley Hauerwas (1940–)

Ethics

How they are often manifested?

Intuitionism presumes human beings have an intuitive moral sense that enables them to detect real moral truths.

Consequentialism bases morality on the consequences of human actions and not on the actions themselves.

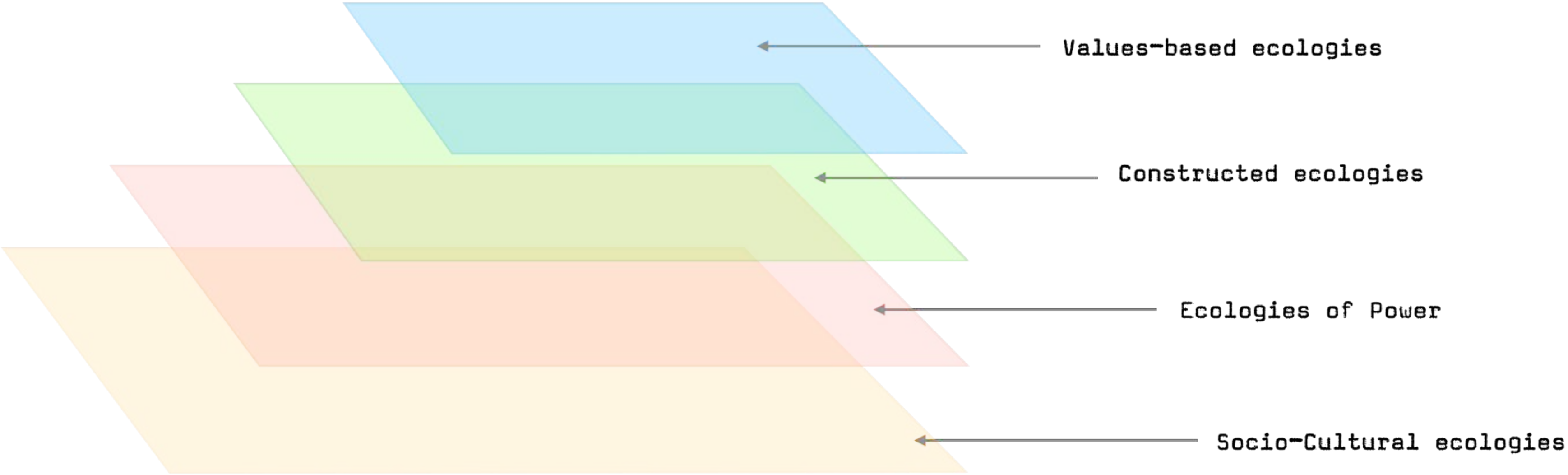
Non-consequentialism is concerned with the actions themselves and not with the consequences.

Virtue ethics is concerned with the way individuals live their lives, and less concerned in assessing particular actions.

Situational ethics argues that individual ethical decisions should be made according to the unique situation rather than prescriptive rules.

Aalto University
School of Science

Ethics & Values across Ecologies in Society



Sawhney, N., and Tran, A., 2020. Ecologies of Contestation in Participatory Design.
In *Proceedings of the 16th Participatory Design Conference (PDC 2020)*, Manizales, Columbia. ACM.

Ethics

Some sources

Ethics Defined, Laura Anabelle, Medium, March 5, 2017.

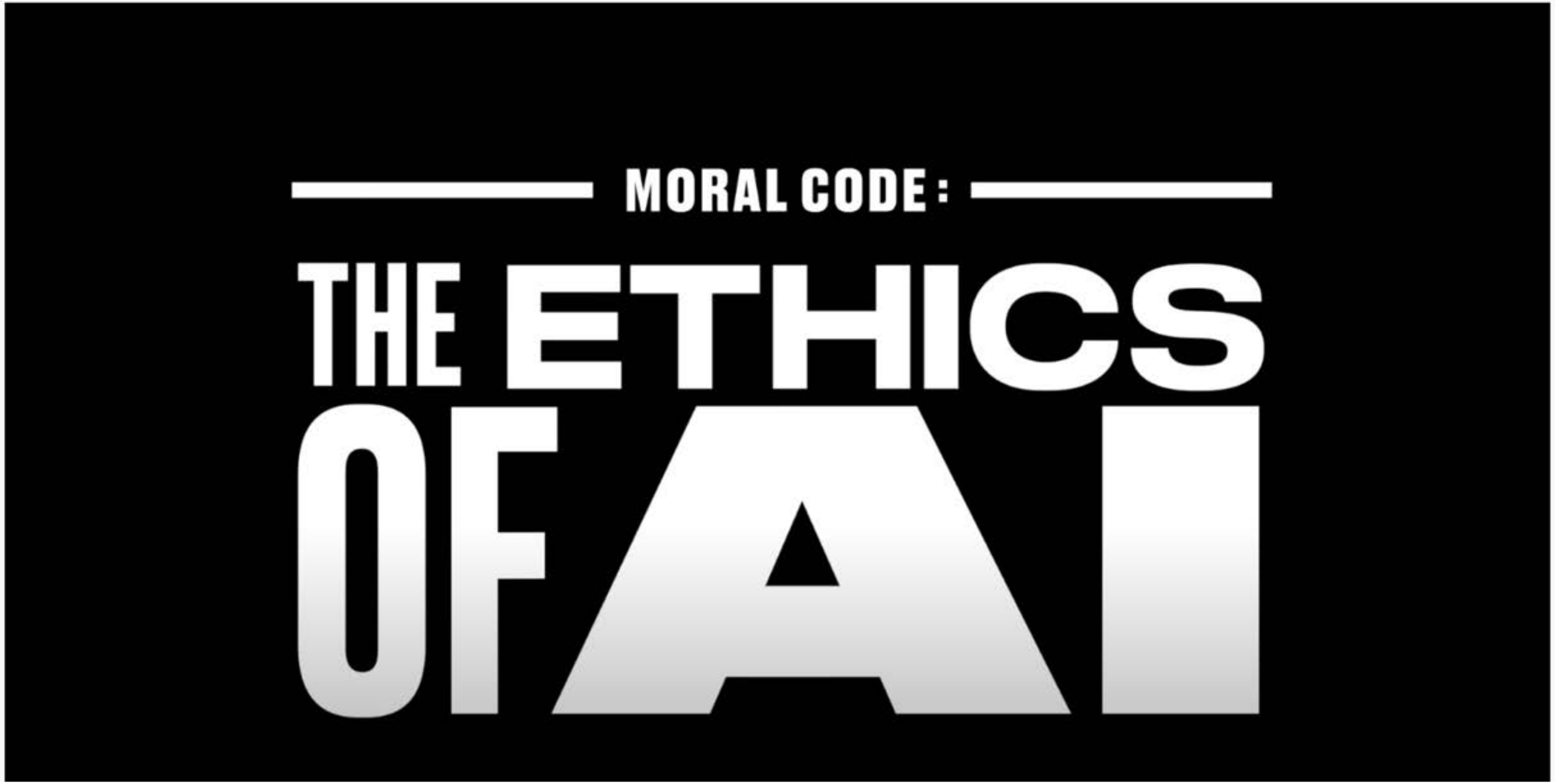
Ethics: A General Introduction, BBC, 2014.

The Hitchhiker's Guide to AI Ethics, B Nalini, Medium, May 1, 2019.

Aalto University
School of Science

Examining Ethics in AI

Aalto University
School of Science

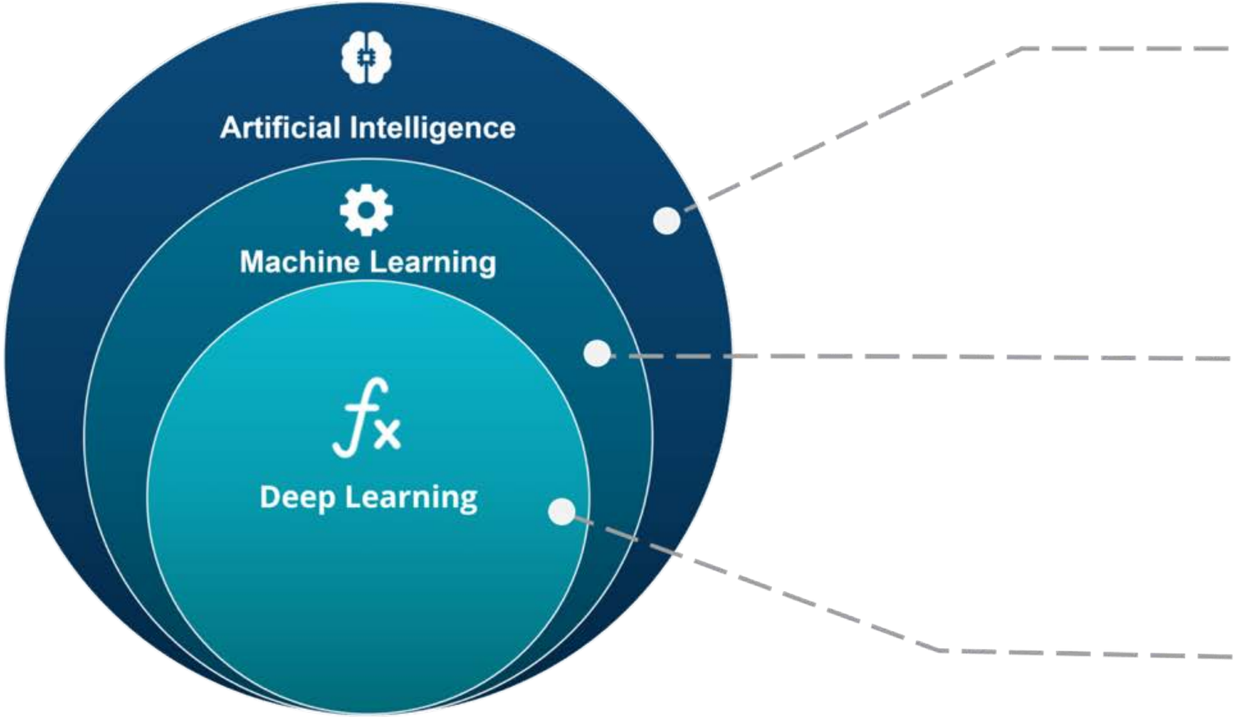


MORAL CODE:

THE ETHICS OF AI

The Atlantic Re:think and Hewlett Packard Labs, June 2018

https://www.theatlantic.com/sponsored/hpe-2018/the-ethics-of-ai/1865/



ARTIFICIAL INTELLIGENCE

A technique which enables machines to mimic human behaviour

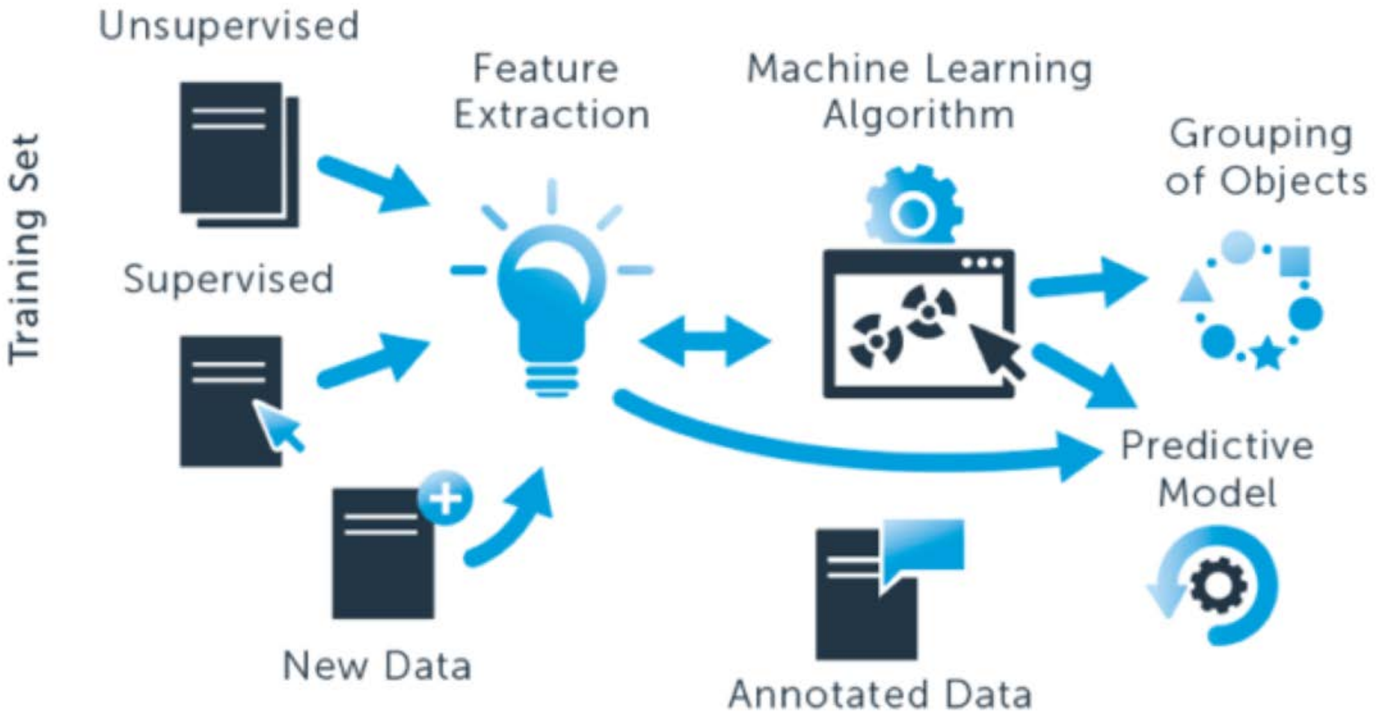
MACHINE LEARNING

Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING

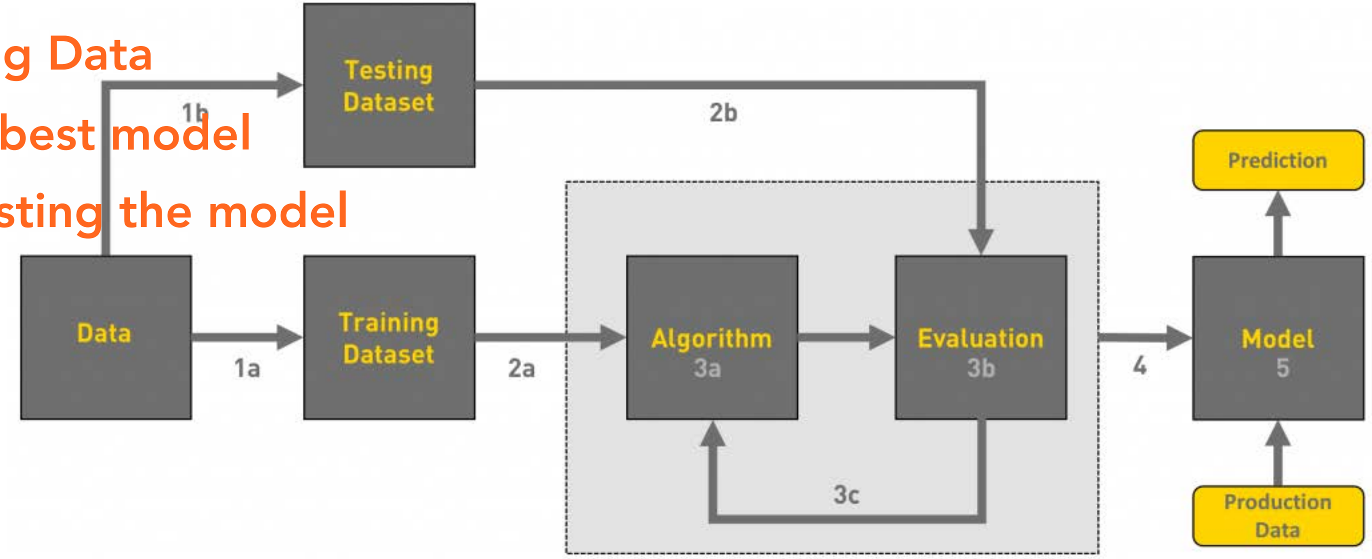
Subset of ML which make the computation of multi-layer neural network feasible

Machine Learning



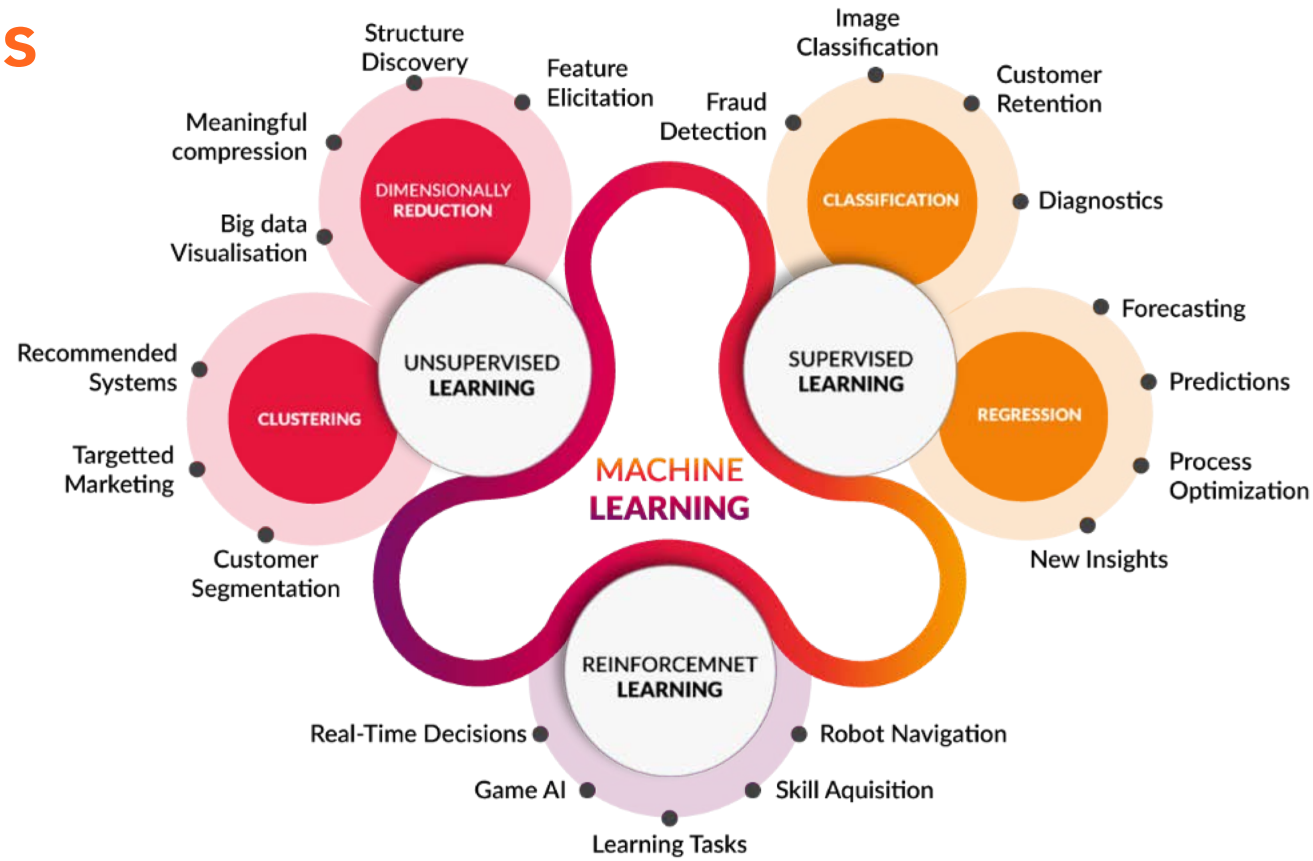
Machine Learning Workflow

1. Gathering Data
2. Pre-processing Data
3. Devising the best model
4. Training & Testing the model
5. Evaluating



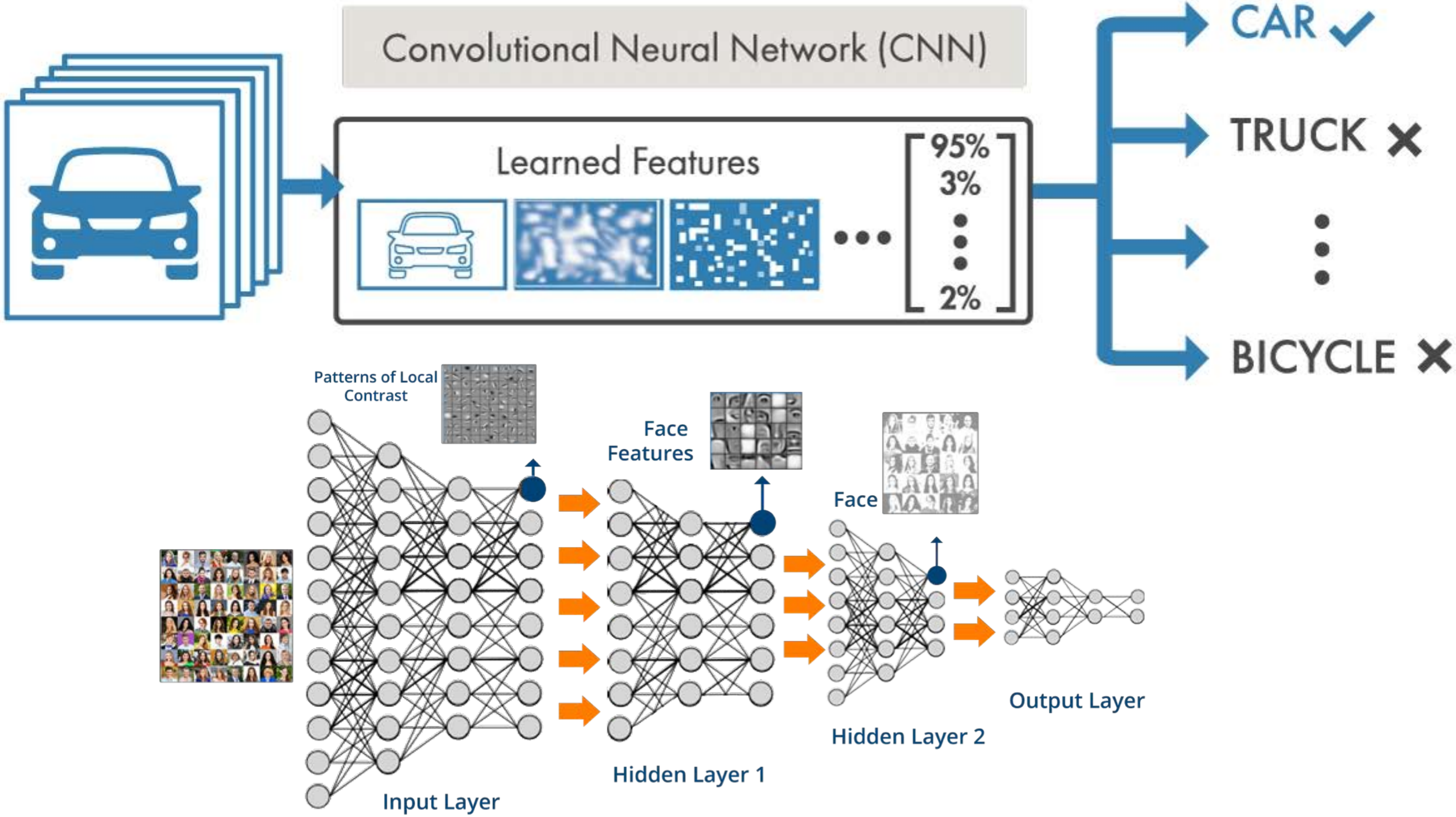
Ayush Pant, Workflow of a Machine Learning project, Towards Data Science, Jan 11, 2019

Machine Learning Models



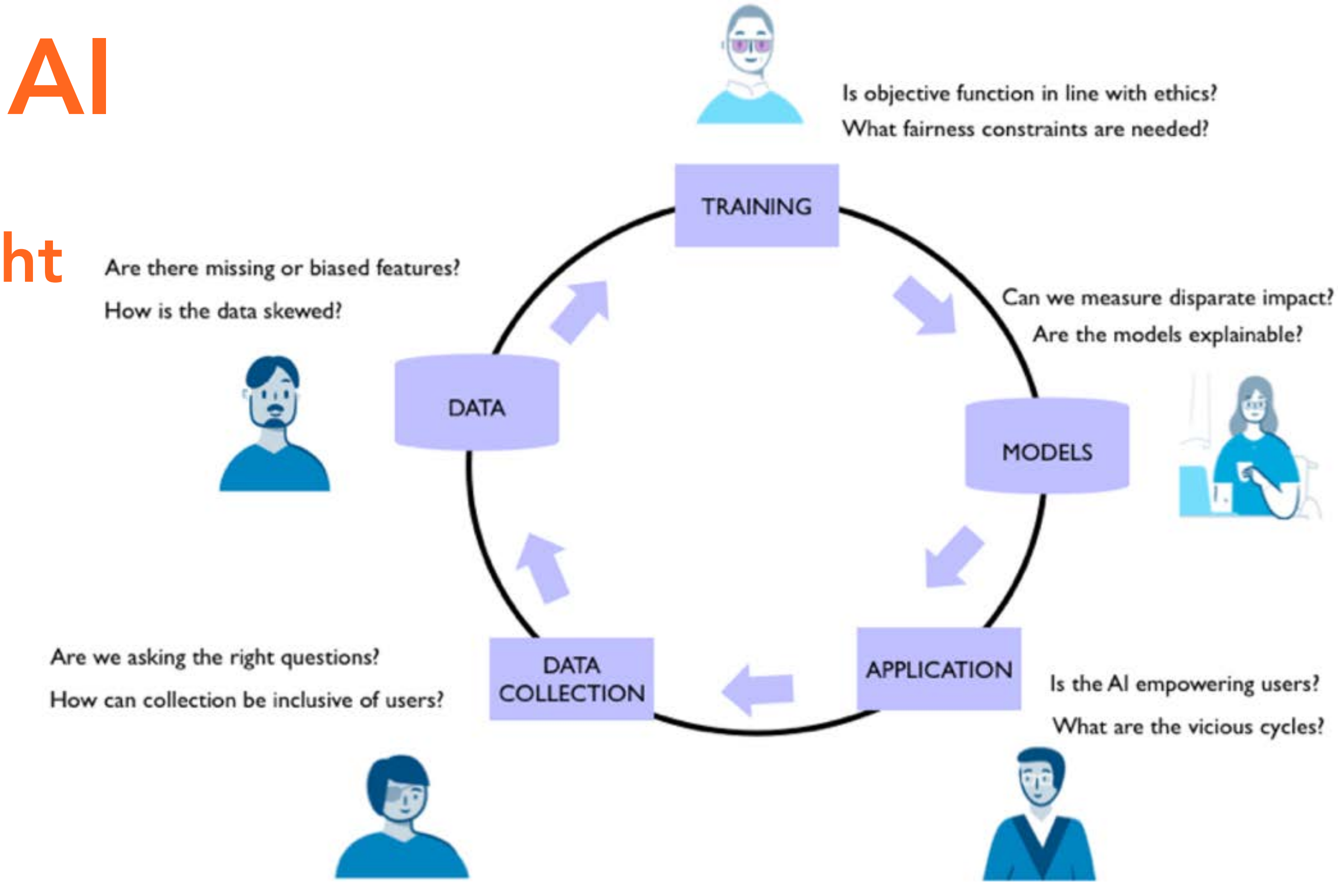
Aalto University School of Science

DEEP LEARNING



Ethics in AI

Asking the right Questions


Is objective function in line with ethics?
What fairness constraints are needed?
TRAINING
Are there missing or biased features?
How is the data skewed?
Can we measure disparate impact?
Are the models explainable?
DATA
MODELS
Are we asking the right questions?
How can collection be inclusive of users?
DATA COLLECTION
APPLICATION
Is the AI empowering users?
What are the vicious cycles?

Charles Earl on Discriminatory Artificial Intelligence, October 4, 2017.

Ethics in AI

Quality of Data, Models, Predictions & Oversight

Messy real-world data with missing, inconsistent and noisy data (outliers) often collected due to human errors or poor understanding of domain.

Incomplete and ambiguous models that provide insufficient coverage or explainability.

Insufficient evaluation and oversight of how the models and predictions are used to influence action.

Quality of AI prediction and outcomes affects the Ethical Quality of it’s impact on humans.

Aalto University
School of Science

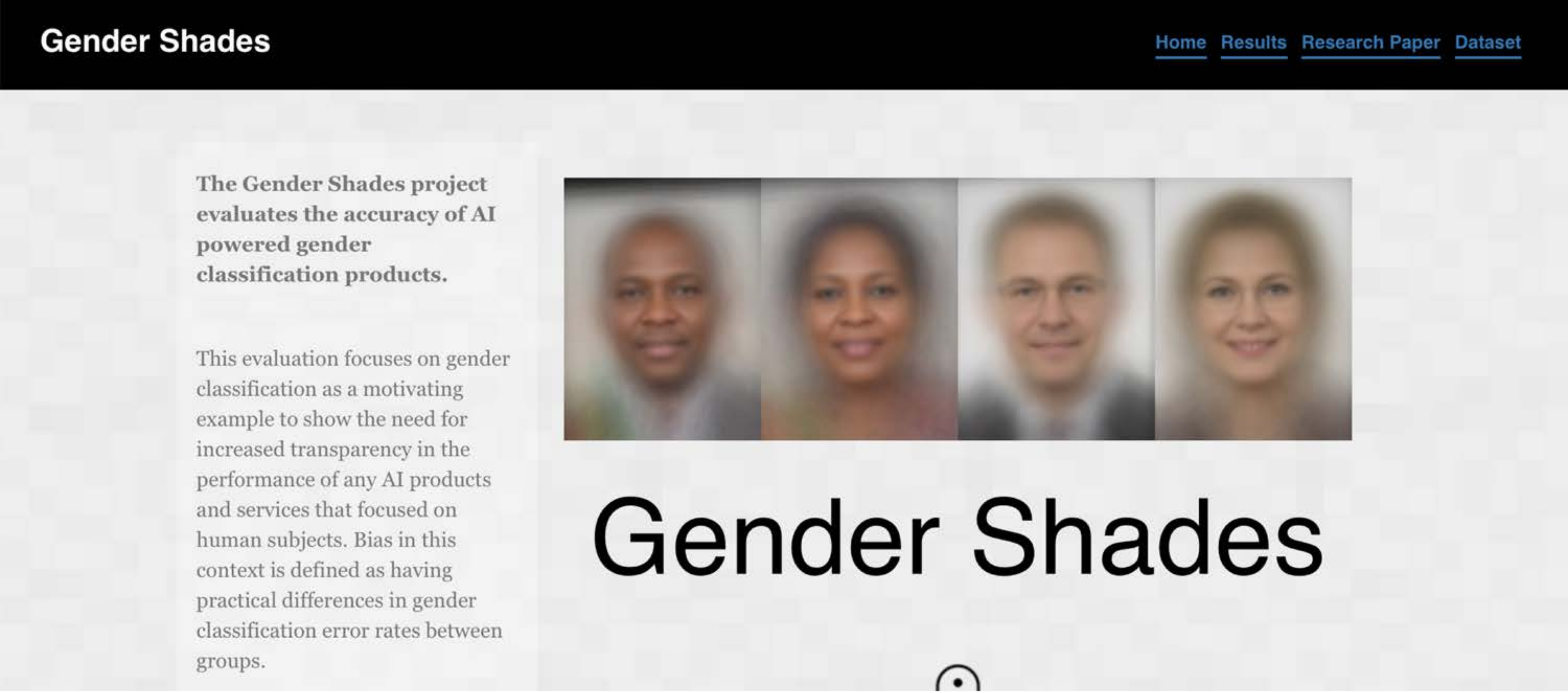
Ethics in AI

Key Concepts

1. Bias and Fairness
2. Accountability and Remediability
3. Transparency, Explainability and Trust
4. Safety and Privacy
5. Value-Alignment

Bias and Fairness

- Cognitive biases are inherent in human decision making and AI can amplify these human biases (scaling them more widely).
- Sources of bias in data include incomplete, skewed and non-representative data used to train machine learning models.
- Machine learning models can also reflect undue prejudice of humans and their flawed social and cultural assumptions.
- Biased algorithmic systems can lead to unfair outcomes, discrimination, and injustice.

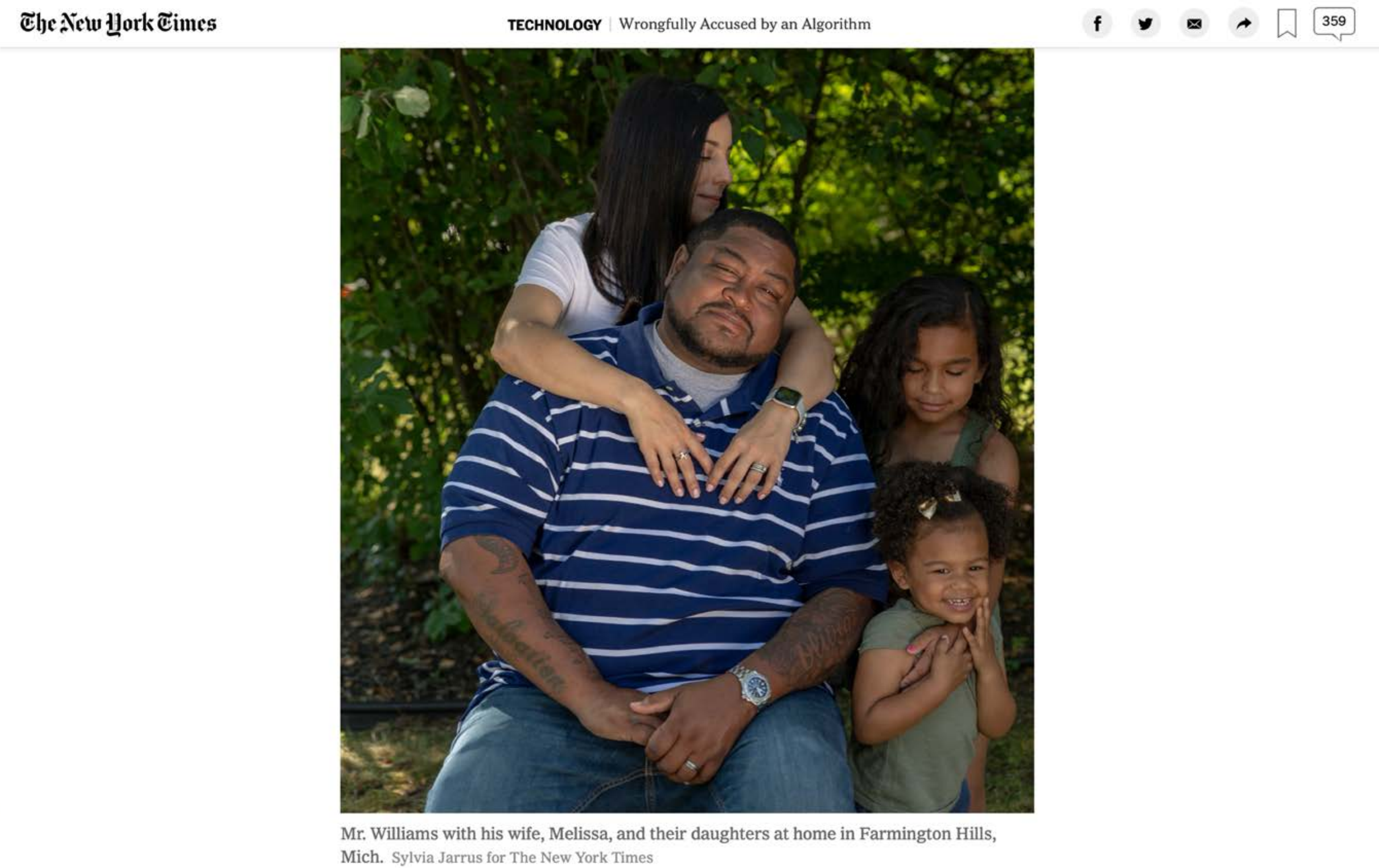


The Gender Shades project evaluates the accuracy of AI powered gender classification products.

This evaluation focuses on gender classification as a motivating example to show the need for increased transparency in the performance of any AI products and services that focused on human subjects. Bias in this context is defined as having practical differences in gender classification error rates between groups.

Gender Shades

http://gendershades.org



The New York Times

TECHNOLOGY | Wrongfully Accused by an Algorithm

Mr. Williams with his wife, Melissa, and their daughters at home in Farmington Hills, Mich. Sylvia Jarrus for The New York Times

Accountability and Remediability

- Accountability includes an obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms.
- Accountability may be achieved by human audits, impact assessments or via governance through policy, regulation or “humans in the loop”.
- Remediation is the process by which unfair or discriminatory practices can be identified and systems modified or withdrawn.

Principles for Accountable Algorithms and a Social Impact Statement for Algorithms.

Transparency, Explainability and Trust

- Transparency seeks to ensure that all human stakeholders can easily understand how an AI system arrives at a decision or recommendation.

- While not all machine learning models are easily interpretable, the goal of explainability is to use models that are inherently explainable and allow humans to trace how decisions are made.

- Improved levels of transparency and explainability enhance the confidence and trust in AI systems.

Aalto University
School of Science

Safety and Privacy

- Safety indicates that AI does not cause accidents or exhibit unintended or harmful behavior.

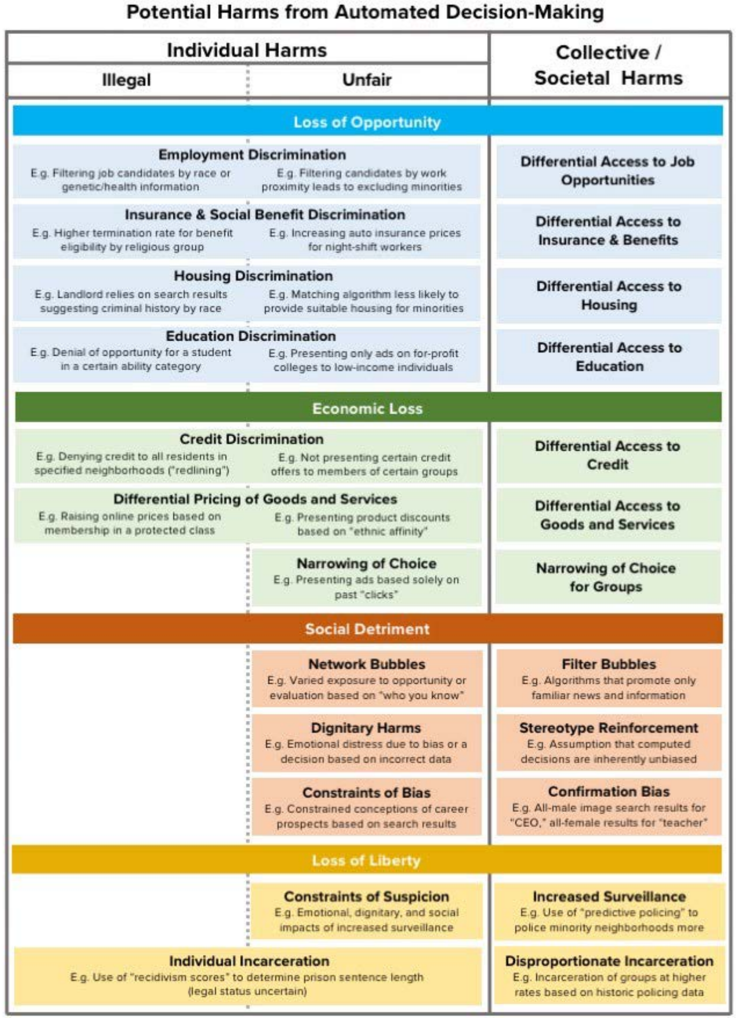
- Privacy suggests that AI must be designed to protect user data and preserve the user's power over access and uses.

- While privacy is a social construct evolving with time and cultural norms, violations can affect human dignity and control.

- In the EU, the General Data Protection Regulation (GDPR) seeks to ensure that systems dealing with user data comply with mandated privacy policies and practices.

Ethics in AI

Potential Harms



Potential Harms from Automated Decision-Making

Individual Harms		Collective / Societal Harms
Illegal	Unfair	
Loss of Opportunity		
Employment Discrimination E.g. Filtering job candidates by race or genetic/health information	E.g. Filtering candidates by work proximity leads to excluding minorities	Differential Access to Job Opportunities
Insurance & Social Benefit Discrimination E.g. Higher termination rate for benefit eligibility by religious group	E.g. Increasing auto insurance prices for night-shift workers	Differential Access to Insurance & Benefits
Housing Discrimination E.g. Landlord relies on search results suggesting criminal history by race	E.g. Matching algorithm less likely to provide suitable housing for minorities	Differential Access to Housing
Education Discrimination E.g. Denial of opportunity for a student in a certain ability category	E.g. Presenting only ads on for-profit colleges to low-income individuals	Differential Access to Education
Economic Loss		
Credit Discrimination E.g. Denying credit to all residents in specified neighborhoods ("redlining")	E.g. Not presenting certain credit offers to members of certain groups	Differential Access to Credit
Differential Pricing of Goods and Services E.g. Raising online prices based on membership in a protected class	E.g. Presenting product discounts based on "ethnic affinity"	Differential Access to Goods and Services
	Narrowing of Choice E.g. Presenting ads based solely on past "clicks"	Narrowing of Choice for Groups
Social Detriment		
	Network Bubbles E.g. Varied exposure to opportunity or evaluation based on "who you know"	**Filter Bubbles** E.g. Algorithms that promote only familiar news and information
	Dignitary Harms E.g. Emotional distress due to bias or a decision based on incorrect data	**Stereotype Reinforcement** E.g. Assumption that computed decisions are inherently unbiased
	Constraints of Bias E.g. Constrained conceptions of career prospects based on search results	**Confirmation Bias** E.g. All-male image search results for "CEO," all-female results for "teacher"
Loss of Liberty		
	Constraints of Suspicion E.g. Emotional, dignitary, and social impacts of increased surveillance	**Increased Surveillance** E.g. Use of "predictive policing" to police minority neighborhoods more
Individual Incarceration E.g. Use of "recidivism scores" to determine prison sentence length (legal status uncertain)		**Disproportionate Incarceration** E.g. Incarceration of groups at higher rates based on historic policing data

FUTURE OF PRIVACY FORUM

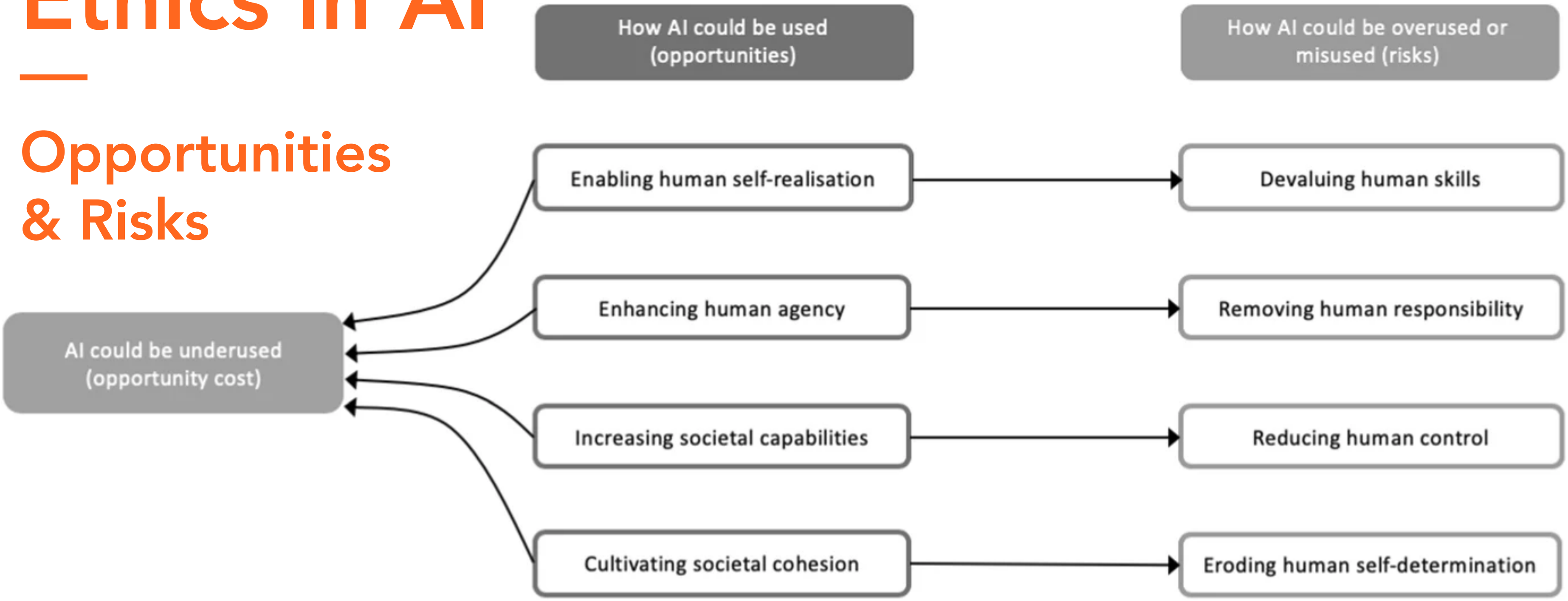
Value-Alignment

- Value-Alignment is a key theme supporting Safety in AI systems.
- It presumes that AI systems should be designed to align with the norms and values of its users.
- Value-Alignment seeks to design methods to prevent AI systems from inadvertently acting in ways inimical to human values.
- The challenge is how AI systems can resolve conflicting norms and values emerging among users; whose values should it privilege at any given time or context?

Aalto University
School of Science

Ethics in AI

Opportunities & Risks



How AI could be used (opportunities)

How AI could be overused or misused (risks)

AI could be underused (opportunity cost)

Enabling human self-realisation → Devaluing human skills

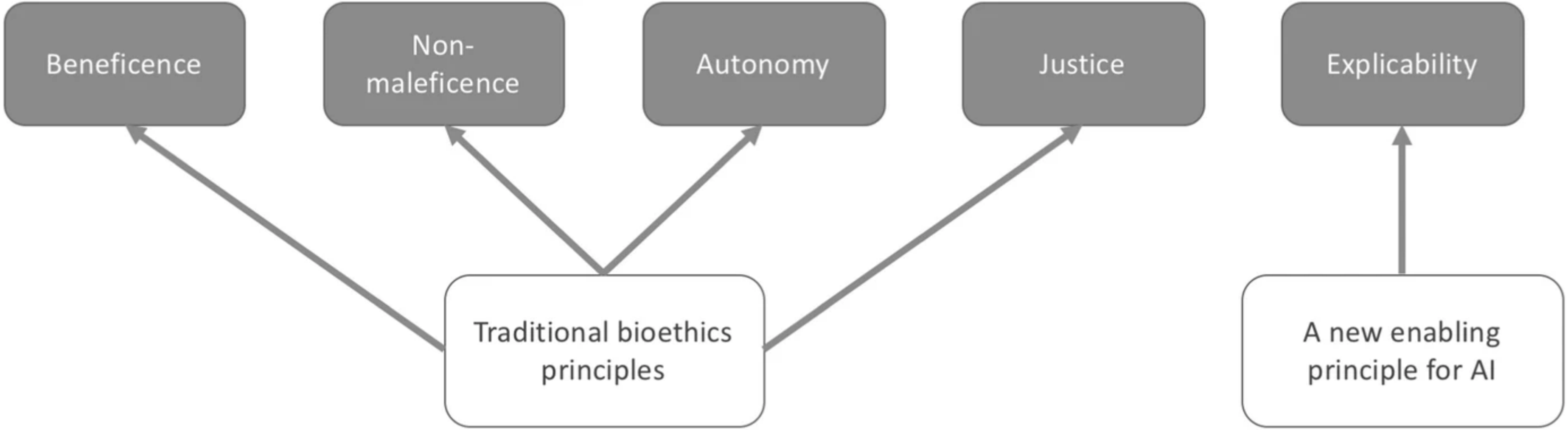
Enhancing human agency → Removing human responsibility

Increasing societal capabilities → Reducing human control

Cultivating societal cohesion → Eroding human self-determination

Luciano Floridi et al, AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Mind and Machines 28, Springer, November 2018.

Ethical Framework for AI extending Bioethics



Beneficence

Non-maleficence

Autonomy

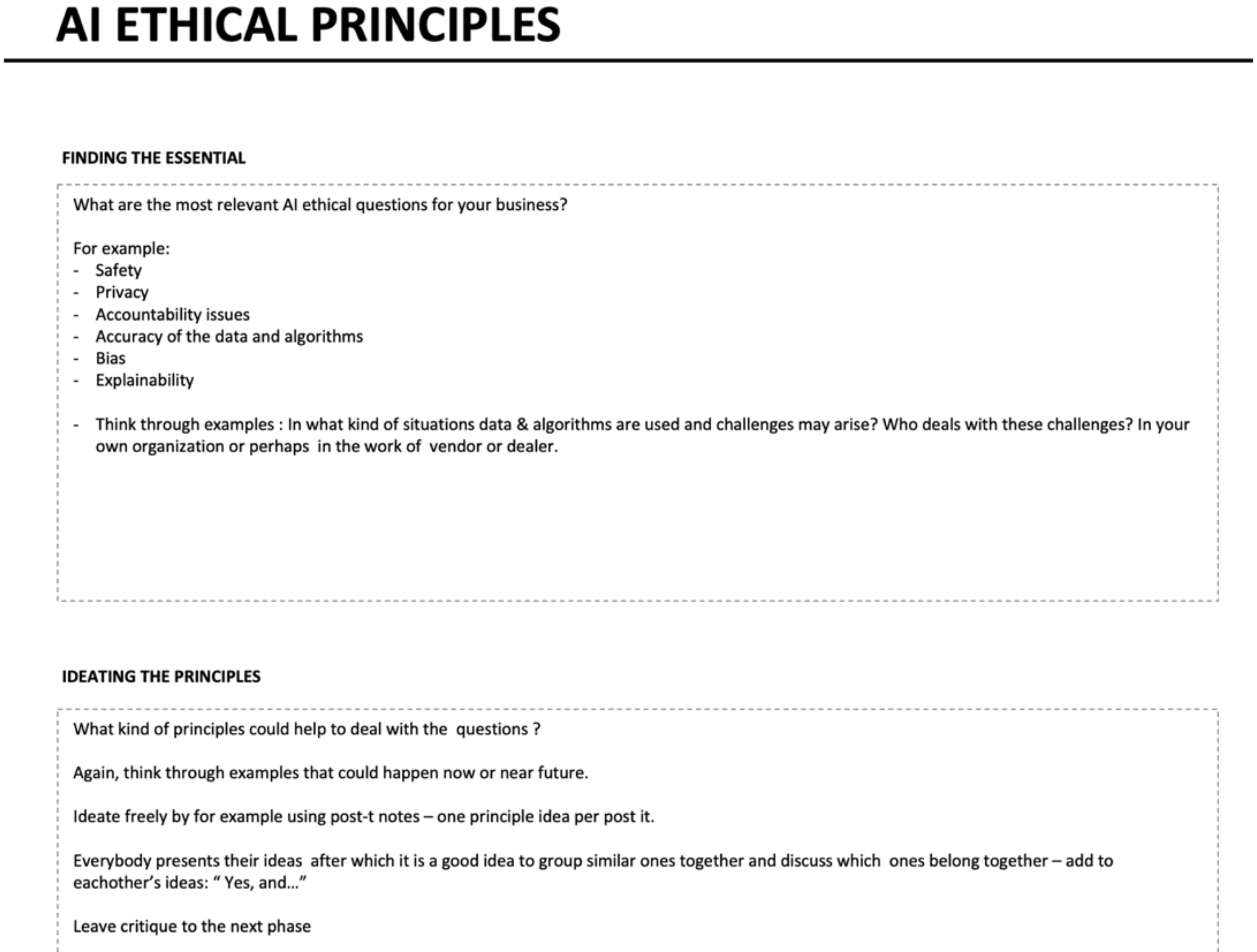
Justice

Explicability

Traditional bioethics principles

A new enabling principle for AI

Luciano Floridi et al, AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Mind and Machines 28, Springer, November 2018.



AI ETHICAL PRINCIPLES

FINDING THE ESSENTIAL

What are the most relevant AI ethical questions for your business?

For example:

- Safety
- Privacy
- Accountability issues
- Accuracy of the data and algorithms
- Bias
- Explainability

- Think through examples : In what kind of situations data & algorithms are used and challenges may arise? Who deals with these challenges? In your own organization or perhaps in the work of vendor or dealer.

IDEATING THE PRINCIPLES

What kind of principles could help to deal with the questions ?

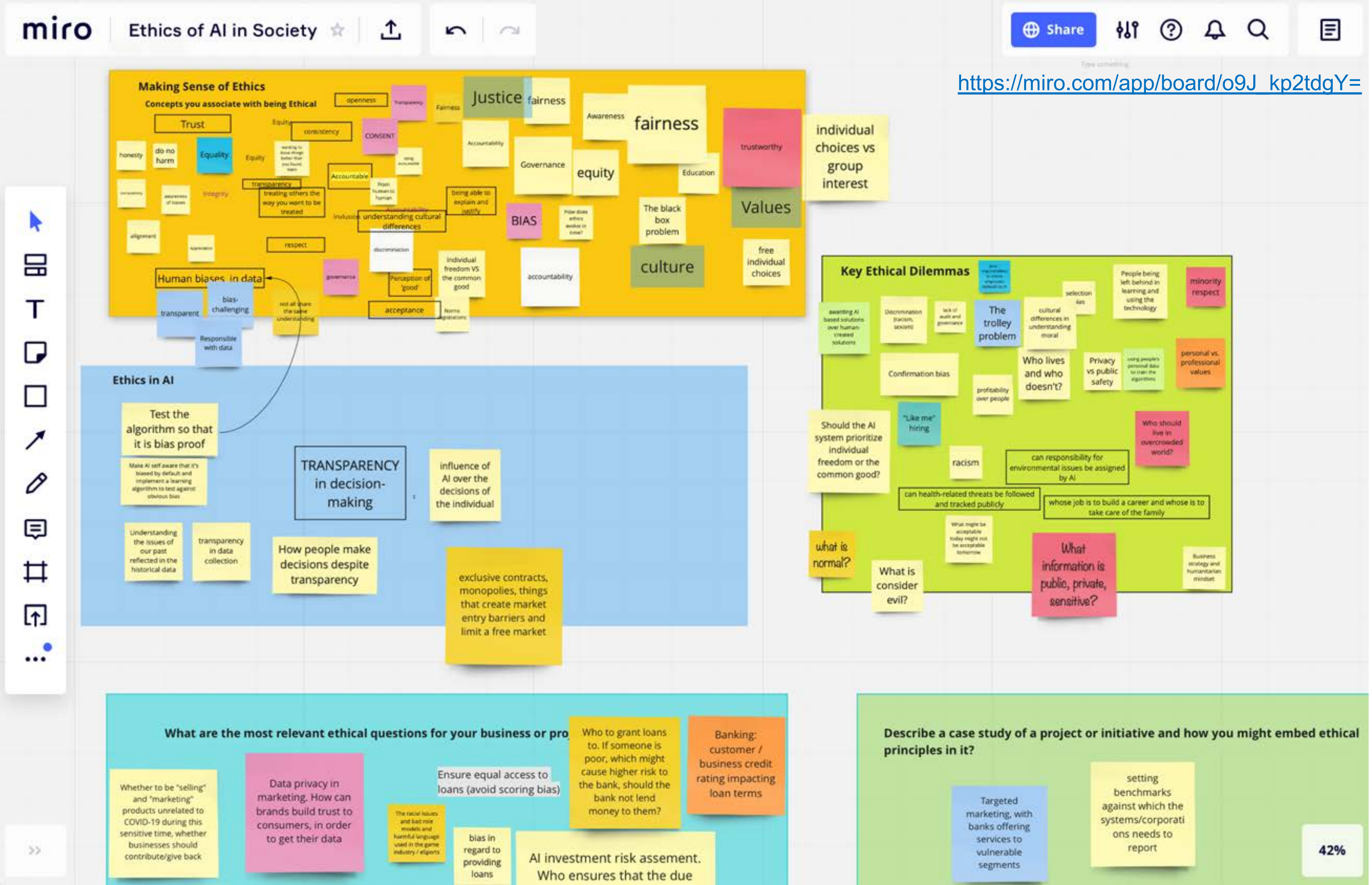
Again, think through examples that could happen now or near future.

Ideate freely by for example using post-t notes – one principle idea per post it.

Everybody presents their ideas after which it is a good idea to group similar ones together and discuss which ones belong together – add to eachother's ideas: “ Yes, and…”

Leave critique to the next phase

https://miro.com/app/board/o9J_kp2tdgY=



Making Sense of Ethics

Concepts you associate with being Ethical

Trust

honesty

do no harm

Equality

Equity

Equity

openness

consistency

CONSENT

Fairness

Accountable

transparency

treating others the way you want to be treated

integrity

From human to human

being able to explain and justify

Inclusion

understanding cultural differences

alignment

respect

discrimination

governance

Perception of 'good'

Individual freedom VS the common good

acceptance

Norms aspirations

Human biases in data

transparent

bias-challenging

Responsible with data

not all share the same understanding

Justice

fairness

Accountability

Governance

equity

Awareness

fairness

Education

trustworthy

BIAS

How does ethics evolve in time?

The black box problem

Values

accountability

culture

free individual choices

individual choices vs group interest

Ethics in AI

Test the algorithm so that it is bias proof

Make AI self aware that it's biased by default and implement a learning algorithm to test against obvious bias

TRANSPARENCY in decision-making

influence of AI over the decisions of the individual

Understanding the issues of our past reflected in the historical data

transparency in data collection

How people make decisions despite transparency

exclusive contracts, monopolies, things that create market entry barriers and limit a free market

Key Ethical Dilemmas

awarding AI based solutions over human-created solutions

Discrimination (racism, sexism)

lack of audit and governance

The trolley problem

cultural differences in understanding moral

selection bias

People being left behind in learning and using the technology

minority respect

Confirmation bias

profitability over people

Who lives and who doesn't?

Privacy vs public safety

using people's personal data to train the algorithms

personal vs. professional values

Should the AI system prioritize individual freedom or the common good?

"Like me" hiring

racism

can responsibility for environmental issues be assigned by AI

Who should live in overcrowded world?

can health-related threats be followed and tracked publicly

whose job is to build a career and whose is to take care of the family

What might be acceptable today might not be acceptable tomorrow

what is normal?

What is consider evil?

What information is public, private, sensitive?

Business strategy and humantarian mindset

What are the most relevant ethical questions for your business or pro

Whether to be "selling" and "marketing" products unrelated to COVID-19 during this sensitive time, whether businesses should contribute/give back

Data privacy in marketing. How can brands build trust to consumers, in order to get their data

The racial issues and bad role models and harmful language used in the game industry / eSports

Ensure equal access to loans (avoid scoring bias)

bias in regard to providing loans

Who to grant loans to. If someone is poor, which might cause higher risk to the bank, should the bank not lend money to them?

Banking: customer / business credit rating impacting loan terms

AI investment risk assement. Who ensures that the due

Describe a case study of a project or initiative and how you might embed ethical principles in it?

Targeted marketing, with banks offering services to vulnerable segments

setting benchmarks against which the systems/corporations needs to report

Aalto University School of Science

AI Ethics in Practice: Designing for Ecosystems

Nitin Sawhney, Ph.D.

Professor of Practice

August 5, 2020

Department of Computer Science

Aalto University
School of Science

Human-Centred Research & Design in Crisis