Automated/Algorithmic Decision Systems – Why Human Autonomy is at Stake

Sabine T. Köszegi
Institute for Management Sciences
Vienna University of Technology

Summerschool
Digital Humanism,
Vienna, September 2022

Outlook

What are Automated Decision Systems (ADS)

Why & where are ADS used?

Which challenges are associated with ADS in the work context?

What needs to be done?

Definitions

What exactly is Artificial Intelligence (AI)?

Artificial intelligence is software for processing structured or unstructured data with four characteristics:
1) it works *autonomously,* i.e. without direct user control,
2) its results are *statistical, i.e. it does* not combine cause and effect,
3) it is *adaptive, i.e. it adjusts* its behaviour as it learns more about the context; and
4) it is *interactive, i.e. it* influences our social and physical environment and vice versa.

Dignum et al. 2020, HLEG 2019

What is an algorithm?

An algorithm is like a recipe, i.e. a prescription for a logical sequence of steps for organising, processing and analysing large amounts of data. Algorithms are the result of modelling, which includes both the formalisation of a problem and a goal.

What are Automated (Algorithmic) Decision Systems (ADS)?

We speak of automated or algorithmic decision systems when algorithms execute decision models and human judgement is replaced in whole or in part by the system.

Anatomy of ADS

Algorithms are the result of modelling, which includes both the formalisation of a problem and a goal:

- Which parameters are relevant for the decision?

- How can these criteria be operationalised?

- Which data are relevant?

- Objective: What specifically is to be optimised? How can this be measured?
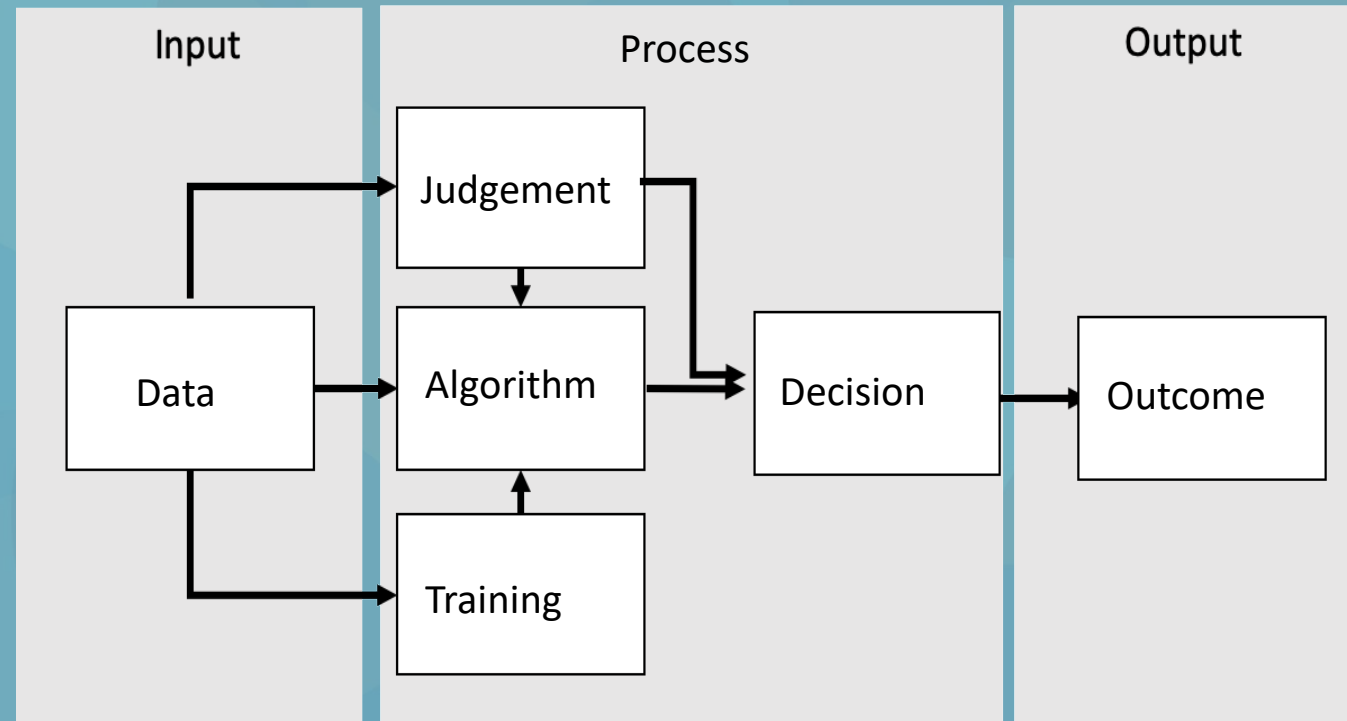
Adaptiert von : Agrawal, Gans & Goldfarb, 2018

Cognitive abilities of AI systems:

Think?
Understand?
Judge?

... Or

Simulate?
Calculate?
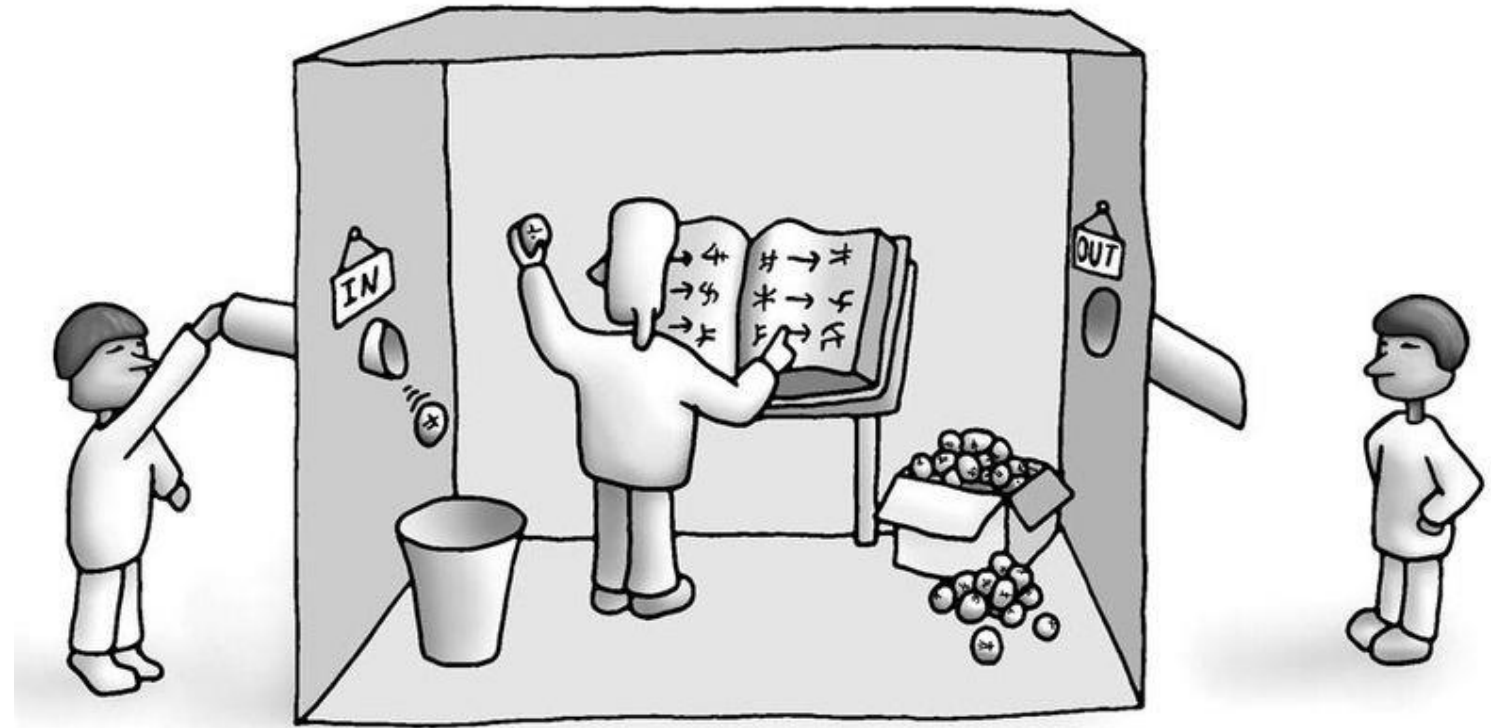
Source: wikicommons

Searl's Chinese Room (1999):
- AI systems are defined purely formally or syntactically (application of rules/algorithms)
- Thinking/understanding/judging/feeling (...) requires consciousness and intentionality
- AI systems do not think/understand/judge/feel (...), but process and simulate

ADS are socio-technical ensembles

https://www.queensu.ca/artsci_online/courses/principles-of-psychology

The term "artificial intelligence" obscures the fact that cultural, social and political values are incorporated into algorithms through the designers' decisions about how to operationalise certain activities, goals or parameters in a model!

Reasons for using Automated Decision Systems?

Reduction of decision alternatives through pre-selection, e.g. recommender systems, spam filters, etc.

Reduction of human error of judgement e.g. driver assistance systems, recruiting, etc.

Reduction of complexity through pattern recognition, e.g. big data analytics, image recognition, etc.

Reduction of uncertainty through prediction e.g. predictive analytics

Increased efficiency through automated processes, e.g. claims handling, etc.

Human error of judgement

Number of alternatives

Complexity

Uncertainty/ risk

Frequency

Challenges (1):
Illusion of more objective and fairer decisions

Example: ADS in recruiting

Expectations

- Management of the alternative space (more suitable candidates and less unsuitable candidates)

- Increasing objectivity
through operationalisation of decision criteria, reduction of bias and increase of fairness

- Increasing the forecasting quality of future performance based on big data (person-job fit)

- Increasing effectiveness through Person-Organisation Fit & Retention

THE HIRING FUNNEL

Sourcing — JOB DESCRIPTIONS ADVERTISING MATCHING HEADHUNTING

Screening — QUALIFICATIONS ASSESSMENTS

Interviewing

Selection — BACKGROUND ✓ OFFER

Evaluation

Source: Bogen et al (2018): Help Wanted, p. 13

Cognitive Bias

What Should We Remember?

We store memories differently based on how they were experienced

We reduce events and lists to their key elements

We discard specifics to form generalities

We edit and reinforce some memories after the fact

We favor simple-looking options and complete information over complex, ambiguous options

To avoid mistakes, we aim to preserve autonomy and group status, and avoid irreversible decisions

To get things done, we tend to complete things we've invested time & energy in

To stay focused, we favor the immediate, relatable thing in front of us

We Need To Act Fast

To act, we must be confident we can make an impact and feel what we do is important

We project our current mindset and assumptions onto the past and future

We think we know what other people are thinking

We simplify probabilities and numbers to make them easier to think about

Not Enough Meaning

We imagine things and people we're familiar with or fond of as better

We fill in characteristics from stereotypes, generalities, and prior histories

We tend to find stories and patterns even when looking at sparse data

We notice flaws in others more easily than we notice flaws in ourselves

We are drawn to details that confirm our own existing beliefs
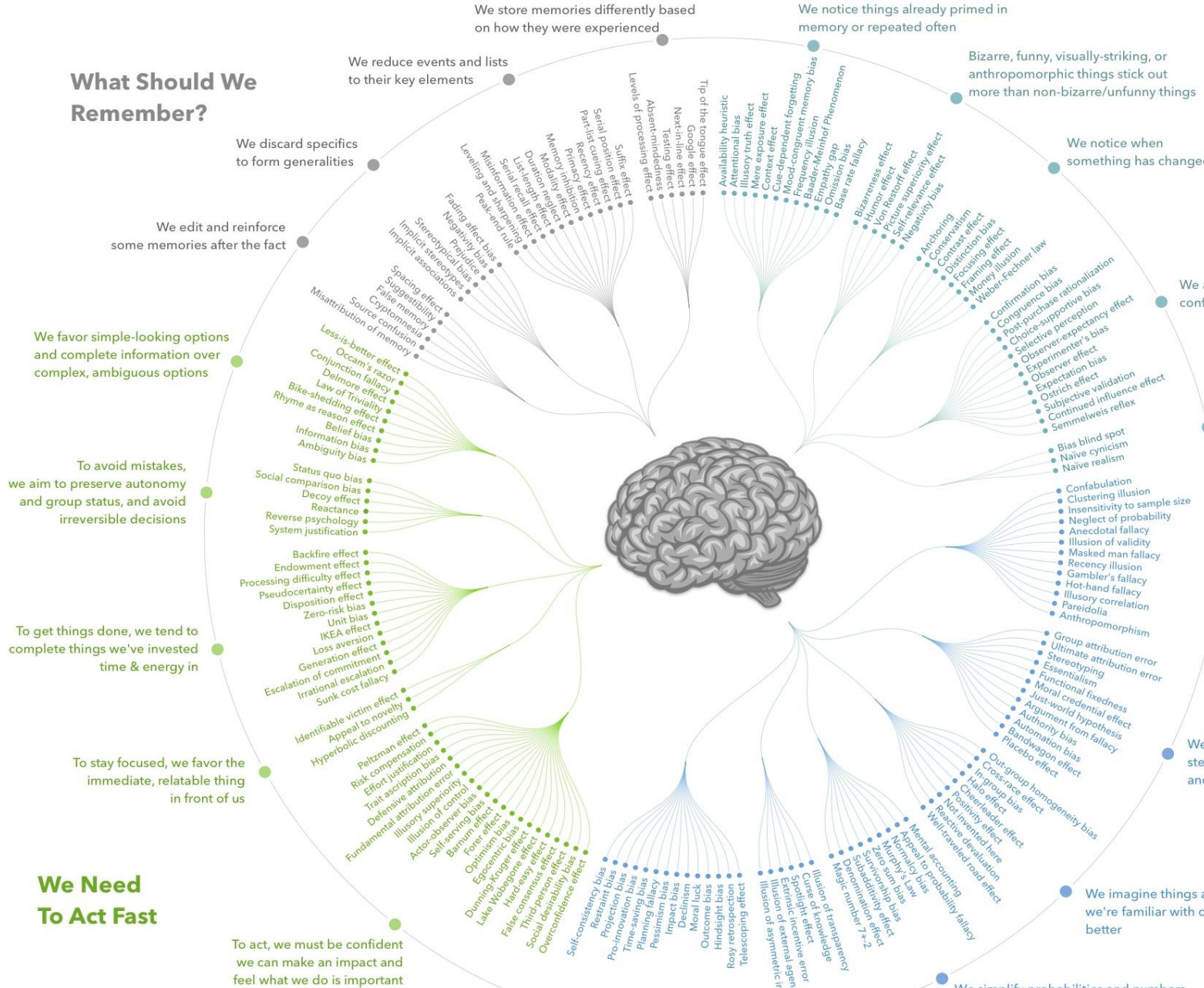
We notice when something has changed

Too Much Information

We notice things already primed in memory or repeated often

Bizarre, funny, visually-striking, or anthropomorphic things stick out more than non-bizarre/unfunny things

Biases (clockwise from top)

Levels of processing effect
Absent-mindedness
Testing effect
Next-in-line effect
Google effect
Tip of the tongue effect
Suffix effect
Serial position effect
Part-list cueing effect
Recency effect
Primacy effect
Memory inhibition
Modality effect
Duration neglect
List-length effect
Serial recall effect
Serial position effect
Misinformation and shaping
Leveling and sharpening
Peak-end rule
Fading affect bias
Negativity bias
Prejudice
Stereotypical bias
Implicit stereotypes
Implicit associations
Spacing effect
Suggestibility
False memory
Cryptomnesia
Source confusion
Misattribution of memory
Availability heuristic
Attentional bias
Illusory truth effect
Mere exposure effect
Context effect
Cue-dependent forgetting
Mood-congruent memory bias
Frequency illusion
Baader-Meinhof Phenomenon
Empathy gap
Omission bias
Base rate fallacy
Bizarreness effect
Humor effect
Von Restorff effect
Picture superiority effect
Self-relevance effect
Negativity bias
Anchoring
Conservatism
Contrast effect
Distinction bias
Focusing effect
Money illusion
Framing effect
Weber-Fechner law
Confirmation bias
Congruence bias
Post-purchase rationalization
Choice-supportive bias
Selective perception
Observer-expectancy effect
Experimenter's bias
Observer effect
Expectation bias
Ostrich effect
Subjective validation
Continued influence effect
Semmelweis reflex
Bias blind spot
Naïve cynicism
Naïve realism
Confabulation
Clustering illusion
Insensitivity to sample size
Neglect of probability
Anecdotal fallacy
Illusion of validity
Masked man fallacy
Recency illusion
Gambler's fallacy
Hot-hand fallacy
Illusory correlation
Pareidolia
Anthropomorphism
Group attribution error
Ultimate attribution error
Stereotyping
Essentialism
Functional fixedness
Moral credential effect
Just-world hypothesis
Argument from fallacy
Authority bias
Automation bias
Bandwagon effect
Placebo effect
Out-group homogeneity bias
Cross-race effect
In-group bias
Halo effect
Cheerleader effect
Positivity effect
Not invented here
Reactive devaluation
Well-traveled road effect
Mental accounting
Appeal to probability fallacy
Normalcy bias
Murphy's Law
Zero sum bias
Survivorship bias
Subadditivity effect
Denomination effect
Magic number 7±2
Illusion of transparency
Curse of knowledge
Spotlight effect
Extrinsic incentive error
Illusion of external agency
Illusion of asymmetric insight
Telescoping effect
Rosy retrospection
Hindsight bias
Outcome bias
Moral luck
Declinism
Impact bias
Pessimism bias
Planning fallacy
Time-saving bias
Pro-innovation bias
Projection bias
Restraint bias
Self-consistency bias
Overconfidence effect
Social desirability bias
Third-person effect
False consensus effect
Hard-easy effect
Lake Wobegone effect
Dunning-Kruger effect
Egocentric bias
Optimism bias
Forer effect
Barnum effect
Self-serving bias
Actor-observer bias
Illusory superiority
Fundamental attribution error
Defensive attribution
Trait ascription bias
Effort justification
Risk compensation
Peltzman effect
Hyperbolic discounting
Appeal to novelty
Identifiable victim effect
Sunk cost fallacy
Irrational escalation
Escalation of commitment
Generation effect
Loss aversion
IKEA effect
Unit bias
Zero-risk bias
Disposition effect
Pseudocertainty effect
Processing difficulty effect
Endowment effect
Backfire effect
System justification
Reverse psychology
Reactance
Decoy effect
Social comparison bias
Status quo bias
Ambiguity bias
Information bias
Belief bias
Rhyme as reason effect
Bike-shedding effect
Law of Triviality
Delmore effect
Conjunction fallacy
Occam's razor
Less-is-better effect

Visual & Algorithmic Design: John Manoogian III

Concept & Categorization: Buster Benson

List of 188 Cognitive Biases: Wikipedia

10

designhacks.co

Exemplary ADS applications in recruiting

– Analysis of data from social media to pre-select suitable candidates according to qualifications, values, salary expectations, etc.

– Analysis of application videos and texts for selection into the shortlist

– Analysis of visual material and videos to determine so-called "inner states" such as emotions, motivation, honesty...

– Analysis of behaviour in serious games to predict future performance behaviour

SMILE	100
JOY	99.865
CONTEMPT	0.00
ANGER	0.00
EXPRESSIVENESS	100.00
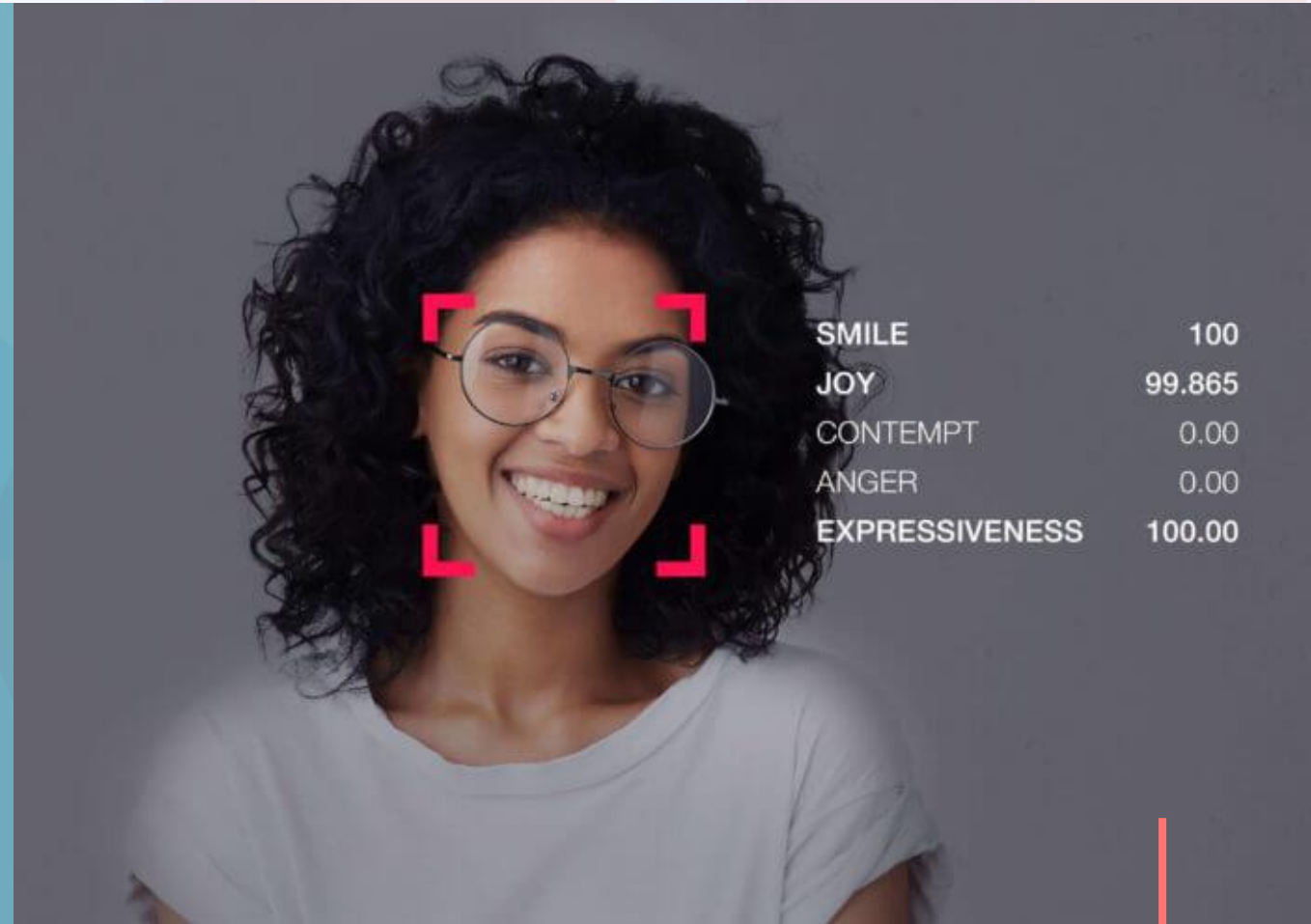
Performance of ADS lags far behind expectations

A recent analysis of 133 AI systems from different industries in Europe shows that **every second system has a gender bias and every fourth system has both a gender & a racial bias**

Source: Smith et al. 2021

Bias in Algorithmic Systems
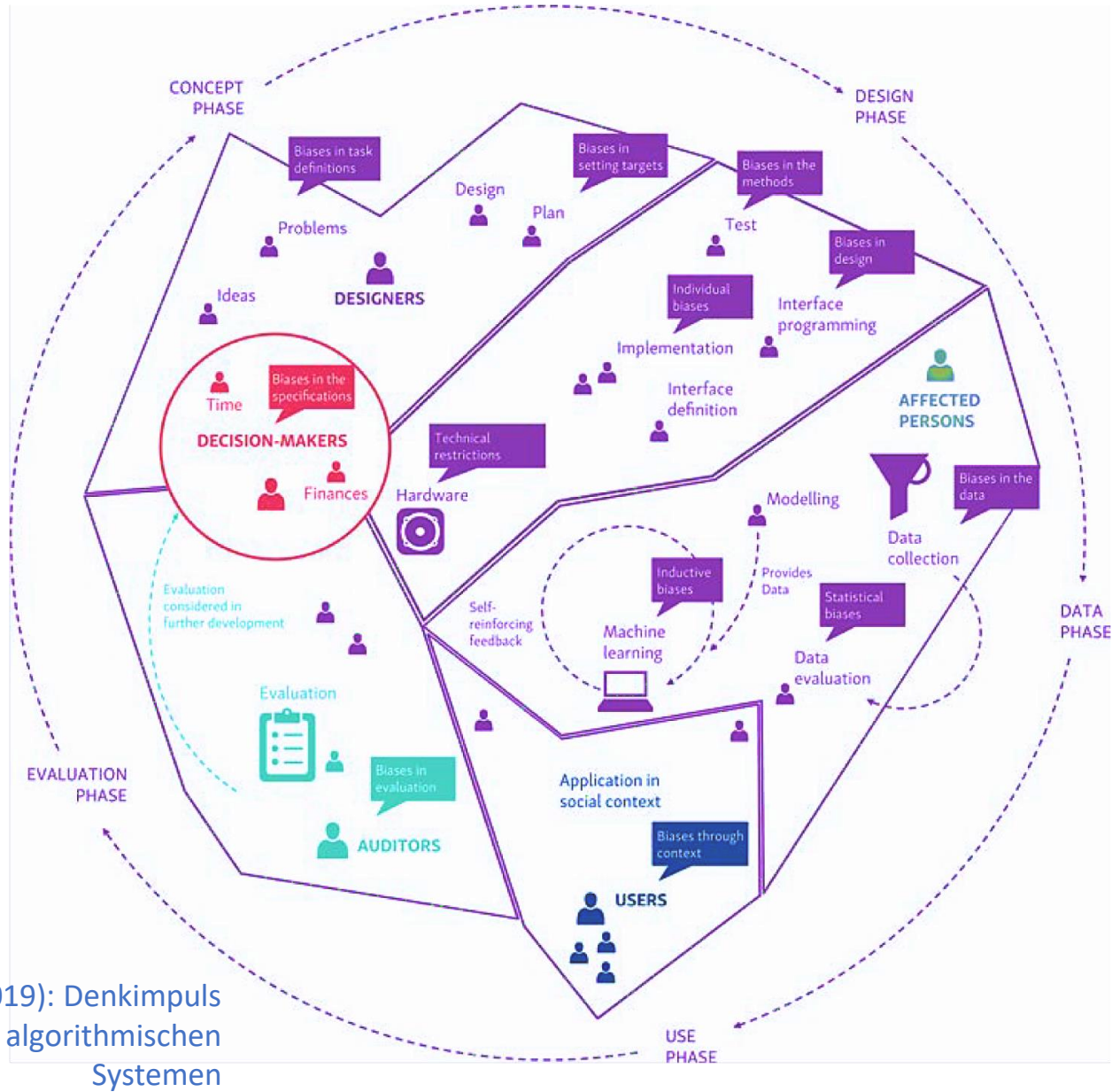
Quelle: Balkow et al. (2019): Denkimpuls Digitale Ethik: Bias in algorithmischen Systemen

Challenges (2): Social Engineering & Filter Bubbels

Nudging

Herding

Conditioning

Eli Pariser 2011: The Filter Bubble,
What the Internet Is Hiding from You

"... First your identity shapes your media (social media), and then they shape what you believe and what matters to you."
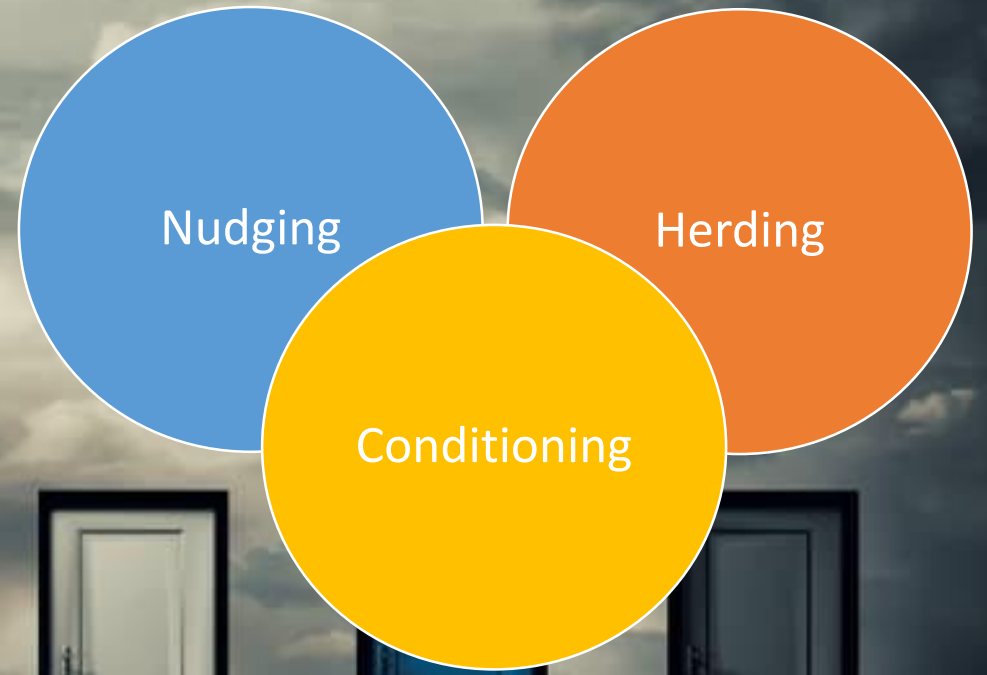
Shoshanna Zuboff (2018, S.335):
Surveillance Capitalism

"The real power of "profiling" user data is to change people's behaviour in the real world."

Challenges (3)
Changing Workers' Role Perceptions, Self Efficacy & Skills

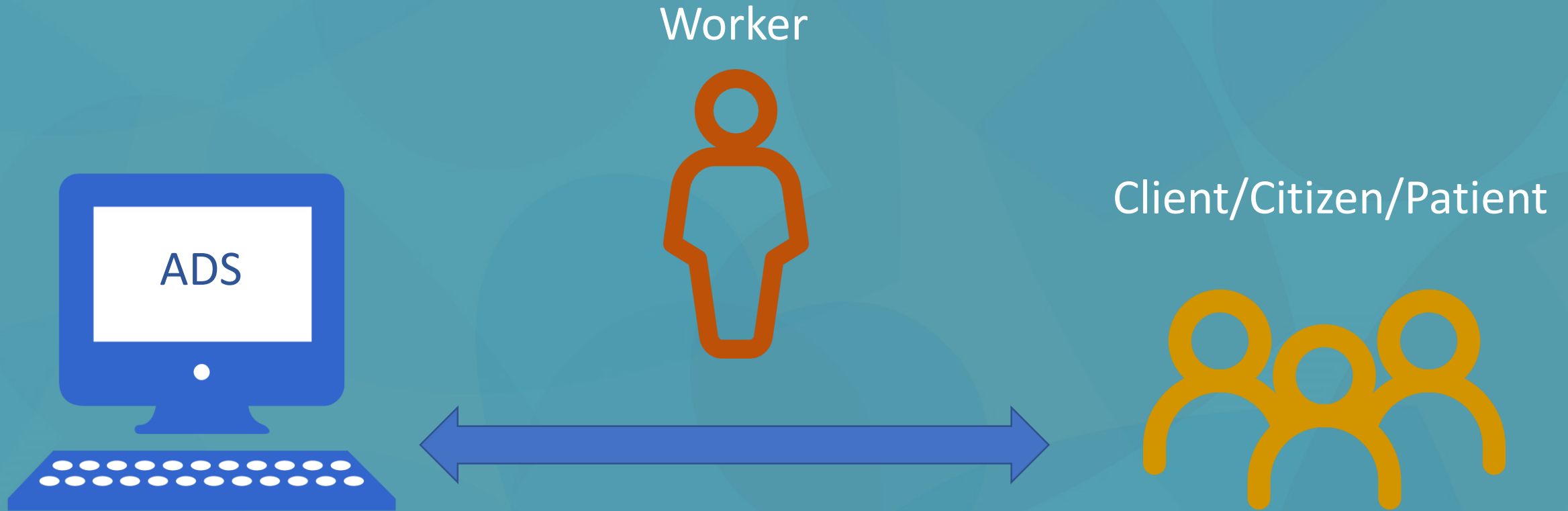# Role Perception

COMPUTER SAYS NO!

Who makes the decision?

Worker

ADS

Client/Citizen/Patient

Who makes the decision?

Worker

ADS

Client/Citizen/Patient

18
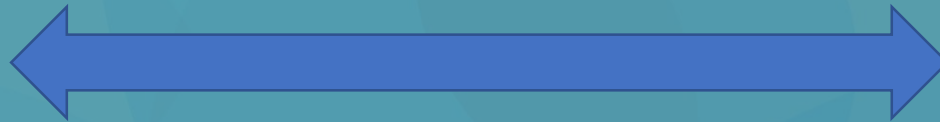
Challenge (4)
Diffusion of Accountability & Lack of Oversight
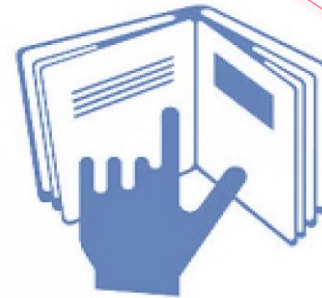
Ironies of Automation

The goal of automation is to replace (unreliable and inefficient) human operators with machines; humans are left with two tasks: Monitoring and intervening in case of an error/problem.

Design error

Irony 1:
Design flaw: only those tasks that can be easily automated are automated!

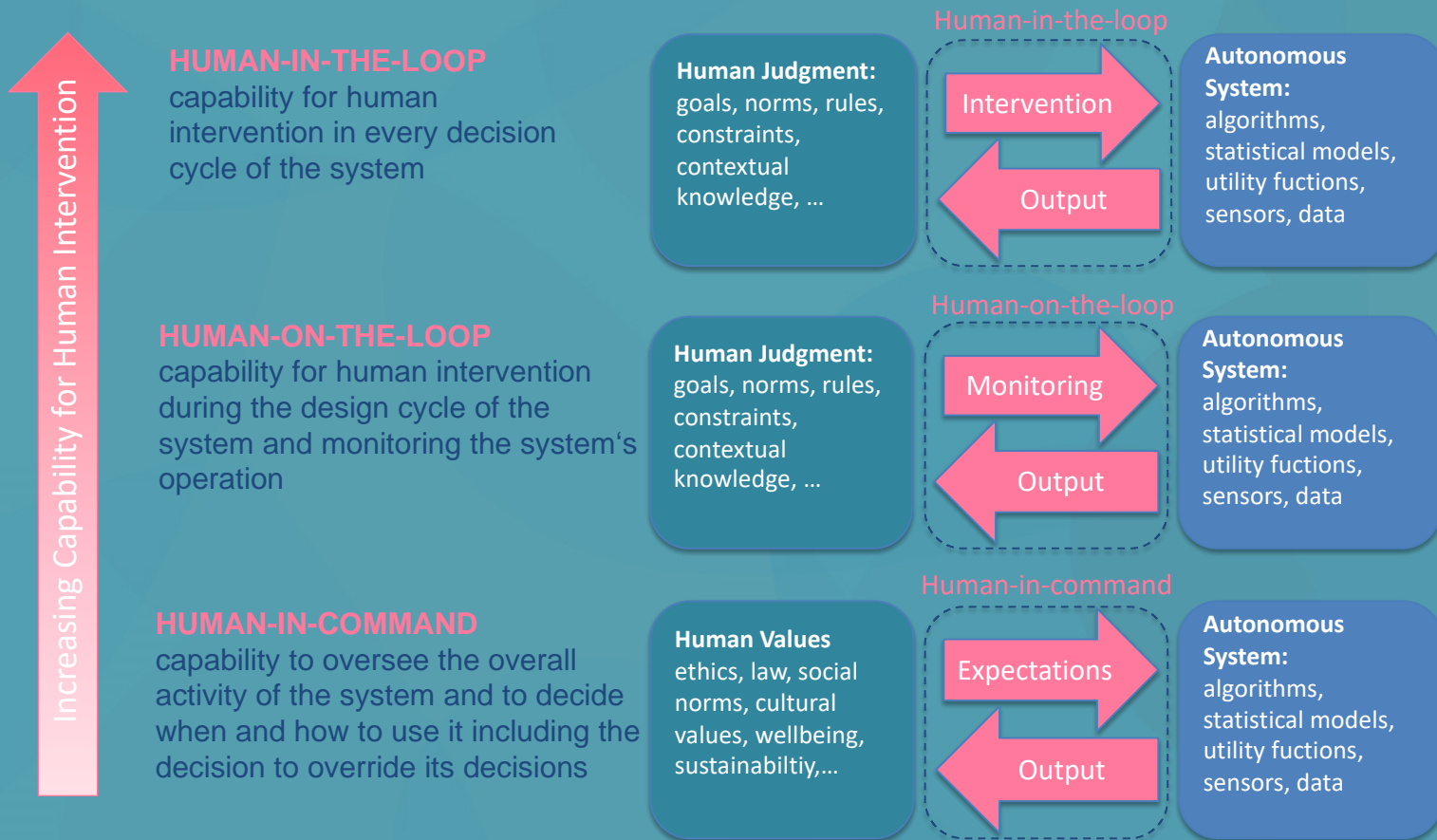### Manual & Cognitive (Control) skills

Irony 2:
Humans are supposed to step in during crises and take over tasks from machines, but skills and knowledge are lost if they are not used regularly!

Monitoring

Irony 3:
Humans are supposed to supervise those machines that have been set up because they can (supposedly) do the job better than humans!

(Lisanne Bainbride, 1986)

Human Agency and Oversight Requires XAI

Increasing Capability for Human Intervention →

HUMAN-IN-THE-LOOP
capability for human intervention in every decision cycle of the system

HUMAN-ON-THE-LOOP
capability for human intervention during the design cycle of the system and monitoring the system's operation

HUMAN-IN-COMMAND
capability to oversee the overall activity of the system and to decide when and how to use it including the decision to override its decisions

Human-in-the-loop

Human Judgment: goals, norms, rules, constraints, contextual knowledge, …

Intervention →

← Output

Autonomous System: algorithms, statistical models, utility fuctions, sensors, data

Human-on-the-loop

Human Judgment: goals, norms, rules, constraints, contextual knowledge, …

Monitoring →

← Output

Autonomous System: algorithms, statistical models, utility fuctions, sensors, data

Human-in-command

Human Values ethics, law, social norms, cultural values, wellbeing, sustainabiltiy,…

Expectations →

← Output

Autonomous System: algorithms, statistical models, utility fuctions, sensors, data

What needs to be done?

Requirements

Input

Privacy and data governance

Process

Technical robustness and safety

Human agency and oversight

Transparency (incl. XAI)

Accountability

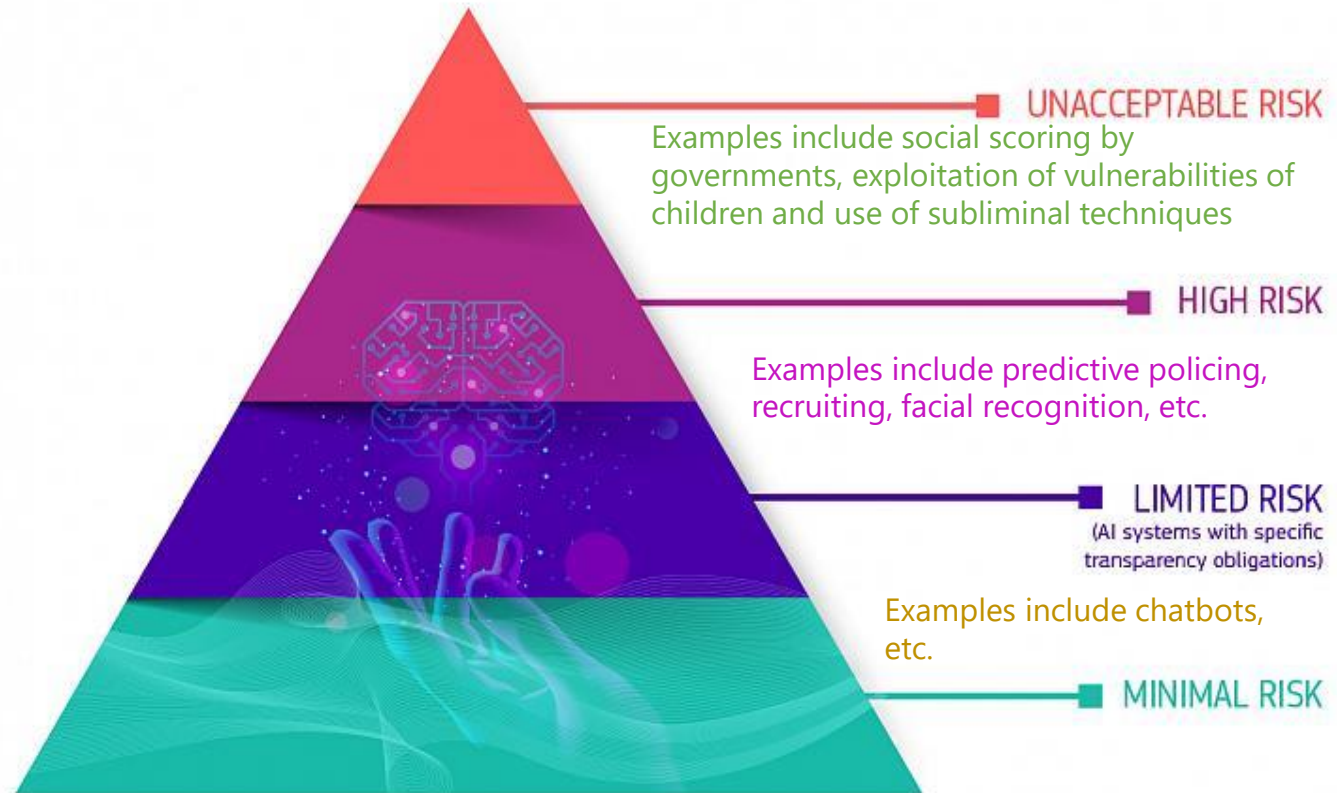**Output**

Diversity, non-discrimination & fairness

Societal & environmental well-being

Trusted AI

HUMAN RIGHTS & RISK ASSESSMENT

UNACCEPTABLE RISK

Examples include social scoring by governments, exploitation of vulnerabilities of children and use of subliminal techniques

HIGH RISK

Examples include predictive policing, recruiting, facial recognition, etc.

LIMITED RISK
(AI systems with specific transparency obligations)

Examples include chatbots, etc.

MINIMAL RISK

Regulation

QUESTIONS?

Sabine T. Köszegi
Institute for Management Sciences
Vienna University of Technology
Sabine.koeszegi@tuwien.ac.at