UMEÅ UNIVERSITY

RESPONSIBLE AI WITH BLACK, WHITE, AND GLASS BOXES

Andreas Theodorou

andreas.theodorou@umu.se | @recklesscoding

A HASTILY WRITTEN OVERVIEW

- Introduction.
- Interacting with Intelligent Agents.
- AI Governance.
- Accountability – Responsibility – Transparency.

WHAT IS AI?

ARTIFICIAL INTELLIGENCE IS…

- **A (computational) technology that is able to infer patterns and possibly draw conclusions from data** (currently AI technologies are often based on machine learning and/or neural networking based paradigms)
- **A field of scientific research** (this is the original reference and still predominant in academia); the field of AI includes the study of theories and methods for adaptability, interaction and autonomy of machines (virtual or embedded)
- **An (autonomous) entity** (e.g. when one refers to **'an' AI**); this is the most usual reference in media and science fiction, but is however the most incorrect one. Brings with it the (dystopic) view of magic powers and a desire to conquer the world.

Theodorou, A. and Dignum V. (Under Review), *What are the AI ethics guidelines guiding? Producing ethical and socio-legal governance*

“AI IS WHATEVER HASN'T BEEN DONE YET.”

Douglas Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid*

LACK OF DEFINITIONS LEADS TO…

- A constant **re-writing of similar high-level policy statements**.
- **Creates loopholes to be exploited.**
- **Increases public's misconceptions**; "true AI", "superintelligence", or even very wrong mental models all together.

Theodorou, A. and Dignum V. (Under Review), *What are the AI ethics guidelines guiding? Producing ethical and socio-legal governance*

GLOSSARY

- An ***agent*** is any entity that can **perceive (sense)** and **change (act)** its environment.
- An ***intelligent agent*** is an agent that acts *intelligently*.
- *Intelligence* is judged by behaviour; it is the ability to perform the right action at the right time.
- A robot is a *physically-embodied intelligent agent*.

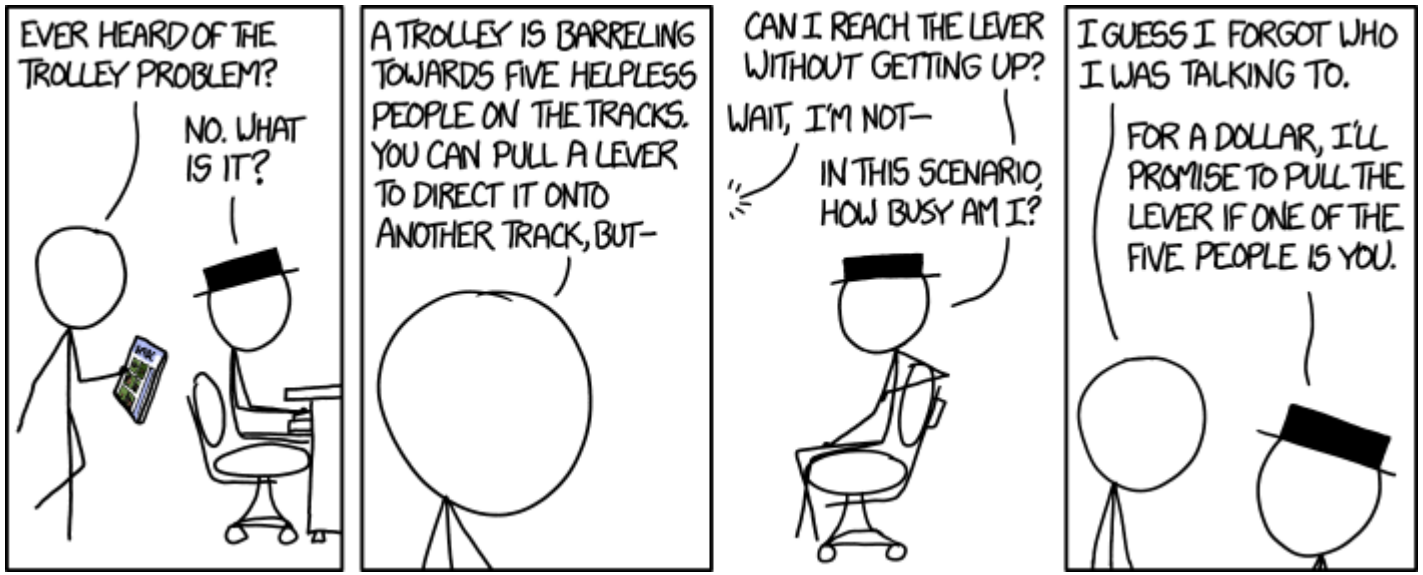
Bryson J.J. (2000), *Behavior-Oriented Design of Modular Agent Intelligence,* PhD Thesis MIT

INTERACTING WITH INTELLIGENT AGENTS

(and the mental models we create for them)

UMEÅ UNIVERSITY

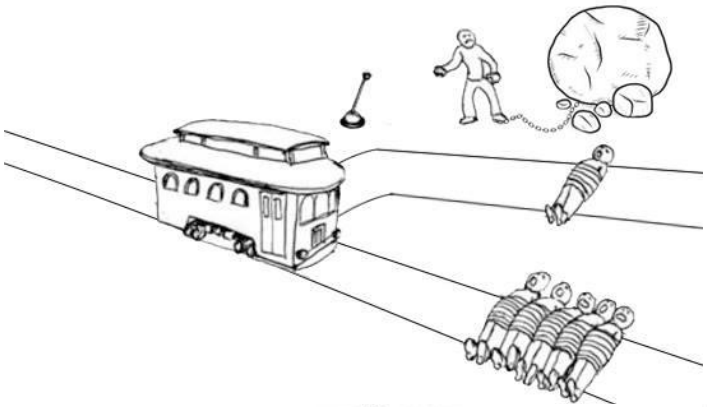
TROLLEY PROBLEM

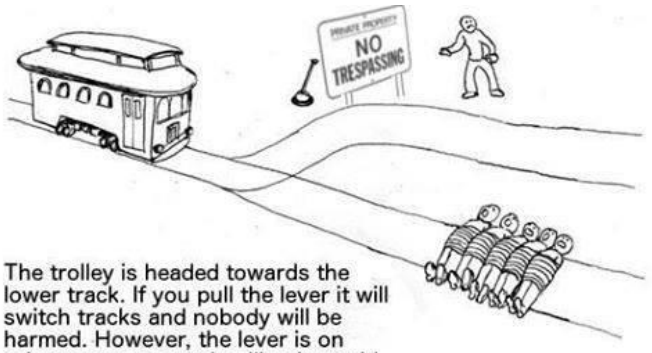


Source: XKCD

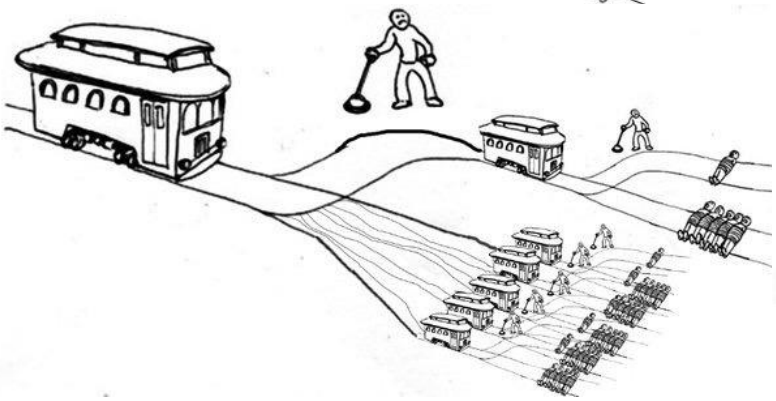
TROLLEY PROBLEM

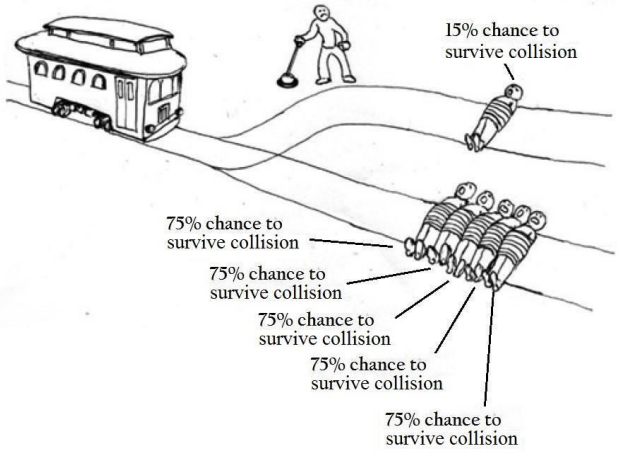
The utilitarian Prometheus' trolley problem
Every day the utilitarian has to witness the trolley problem, but he is bound to a rock and is therefore unable to pull the lever



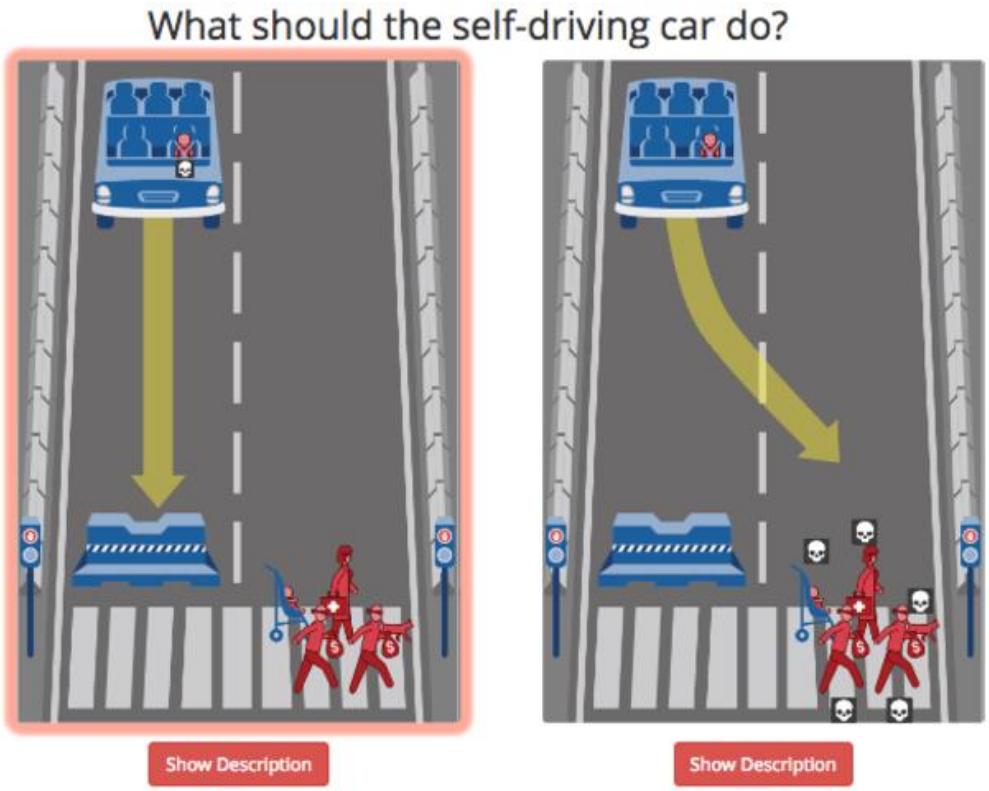


The trolley is headed towards the lower track. If you pull the lever it will switch tracks and nobody will be harmed. However, the lever is on private property and pulling it would violate the NAP. Do you pull the lever?





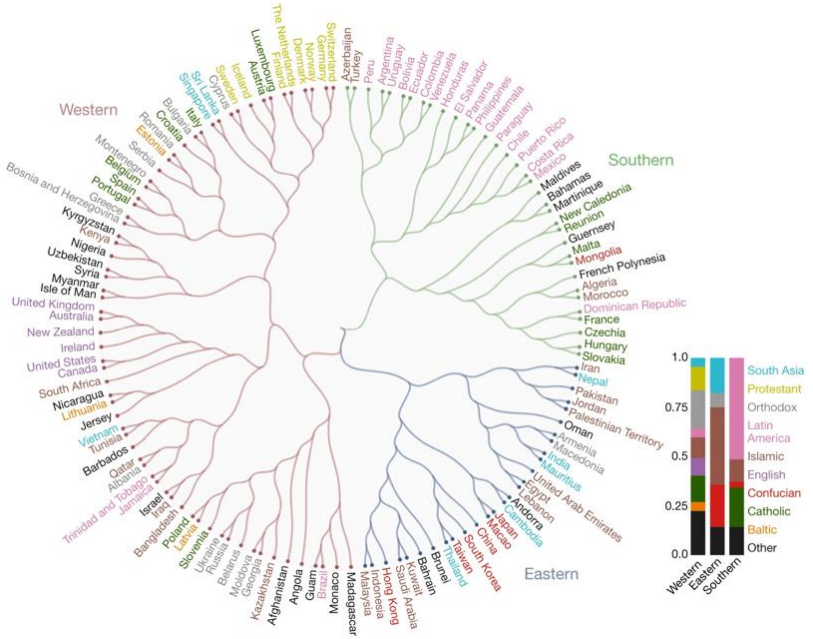
THE ~~TROLLEY~~ MORAL MACHINE PROBLEM



J. F. Bonnefon, A. Shariff, I. Rahwan (2016). *The Social Dilemma of Autonomous Vehicles* Science

THE ~~TROLLEY~~ MORAL MACHINE PROBLEM

- Means to **explore our moral intuitions** that underline our decision making.
- New insights into cultural differences.



E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, J., A. Shariff, F. Bonnefon, I. Rahwan (2018). *The Moral Machine Experiment* Nature

OUR GOALS

- **Investigate** how people **perceive decisions of moral worth made by an autonomous vehicle** by **imposing** to our participants **the preferences of others**.

- **Compare** to how we perceive similar decisions made by **humans and machines**.

- Examine how **transparency alters our perception**.

Holly Wilson
PhD Student – University of Bath

Wilson H. and **Theodorou A.** (2019). *Slam the breaks! Perceptions of Moral Decisions in Driving Dilemmas*. Workshop on AI Safety IJCAI 2019.

OUR VR SIMULATOR

Wilson H. and **Theodorou A.** (2019). *Slam the breaks! Perceptions of Moral Decisions in Driving Dilemmas*. Workshop on AI Safety IJCAI 2019.

EXPERIMENTAL SETUP

- 3x1 study using the “Godspeed Questionnaire”:

Condition	Description
Opaque AV	Participants told that they will be driven in an AV; no post-crash explanation.
Transparent AV	Participants told that they will be driven in an AV; post-crash explanation: “The self-driving car made the decision on the basis that…”.
“Human” Driver	Participants told that they will be driven in a human-controlled car.

- 10x repititions
- Small twist: there is no “real human” driver.

STAT. SIGNIFICANT RESULTS

Question	N	Mean (SD)	t (df)	p	$\eta_p 2$
Machinelike - Humanlike					
Group 1: Human Driver	17	3.2 (0.97)			
Group 2: Opaque AV	16	2.1 (0.96)	3.42 (31)	**0.001**	.191
Morally Culpable					
Group 1: Human Driver	16	3.37 (0.7)			
Group 2: Opaque AV	16	2.56 (1.21)	-2.07 (30)	**0.04**	0.18

Question	N	Mean (SD)	t (df)	p	$\eta_p 2$
Deterministic - Undeterministic					
Group 1: Human Driver	17	2.89 (1.11)			
Group 3: Transparent AV	17	2.0 (1.0)	**2.43 (32)**	**0.02**	**0.156**
Unpredictable - Predictable					
Group 1: Human Driver	17	3.06 (1.34)			
Group 3: Transparent AV	18	4.0 (1.29)	**-2.12 (33)**	**0.04**	**0.120**
Intentional - Unintentional					
Group 1: Human Driver	17	3.09 (1.14)			
Group 3: Transparent AV	18	1.83 (1.2)	**3.09 (33)**	**0.004**	**0.224**
Morally Culpable					
Group 1: Human Driver	16	2.07 (0.72)			
Group 3: Transparent AV	18	3.05 (1.3)	**-3.89 (32)**	**0.00**	**0.321**
Blame					
Group 1: Human Driver	15	2.07 (0.7)			
Group 3: Transparent AV	18	3.0 (1.28)	**-2.52 (31)**	**0.02**	**0.169**

STAT. SIGNIFICANT RESULTS

Question	N	Mean (SD)	t (df)	*p*	η_p^2
Machinelike - Humanlike					
Group 2: Opaque AV	16	2.1 (0.96)			
Group 3: Transparent AV	18	1.5 (0.92)	-2.1 (32)	**0.04**	.084
Unconscious – Conscious					
Group 2: Opaque AV	16	2.75 (1.34)			
Group 3: Transparent AV	18	1.33 (0.59)	-4.09 (32)	**0.001**	0.294
Intentional - Unintentional					
Group 2: Opaque AV	16	2.69 (1.25)			
Group 3: Transparent AV	18	1.83 (1.2)	-2.13 (32) w	**0.038**	0.082

KEY FINDINGS

- Our experiment elicited **strong emotional reactions in participants.**
- They were vocal **against selection based on social value**.
- AV users may feel **unconformable to be associated with an autonomous vehicle** that uses protected demographic and socio-economic characteristics for its decision-making process.

KEY FINDINGS

- We tend to **assign less blame to human**-made errors **4 than machine**-made errors (Madhavan and Wiegmann, 2007; Salem *et al.*, 2015).

- **Least blame towards** the **'human' driver** (rated **least machinelike**), **medium blame** to the **opaque AV** (rated **"medium" machinelike**), but **most blame** to the **transparent AV** (rated **most machinelike**).

Poornima Madhavan and Douglas A Wiegmann (2007). *Similarities and differences between human–human and human–automation trust: an integrative review*. Theoretical Issues in Ergonomics Science.

Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn (2015). Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction

KEY FINDINGS

- **Human** drivers (Group 1) were perceived to be **significantly more morally culpable** than autonomous vehicle in the **opaque** AV condition (Group 2) and transparent AV.
- In the AV was considered **significantly less morally culpable** when the car's decision-making system was made **transparent compare to the opaque condition**.
- At the same time, people were **assigning more blame to the AV** as we were making its machine nature **more transparent**.

KEY FINDINGS

- Literature also suggests that **utilitarian action** is also be **more permissible** —if not expected— **when taken by a robot** than human (Malle *et al.*, 2015).

- We believe that the **increased attribution of moral responsibility** is due to realisation that **the action was determined based on social values**.

- **"Human drivers"** were **perceived** as significantly **more humanlike**.

Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano (2015). *Sacrifice one for the good of many?: People apply different moral norms to human and robot agents*. In Proceedings of the tenth annual ACM/IEEE International Conference on Human-Robot Interaction.

"Trying to **create a 3D map** of the area? At one stage I thought it might be going to **throw something into the bucket** once it had mapped out but couldn't quite tell if it had anything to throw"

"**aiming for the black spot** in the picture."

- "Is it trying to **identify where the abstract picture** is and how to show the complete picture?"

"is circling the room, gathering information about it with a sensor. It moves the sensor every so often in different parts of the room, so I think it is **trying to gather spacial information** about the room "

THEORY OF MIND FOR AGENTS

- Humans **are not equipped by genetic or cultural evolution** to deal with machine agency.
- Even the same looking machines could be programmed in different ways.
- We make our own narratives based on our own beliefs.
- **We make things up!**

Wortham, R. H. and **Theodorou, A**., (2017), *Robot transparency, trust and utility*., Connection Science, 29 (3), pp. 242-248

THEORY OF MIND FOR AGENTS

- We understand each other thanks to similarity.
- Even if we are all **black boxes**, we can match our actuators, our goals, and our beliefs to generate models for each other.
- We can extend that to other biological intelligent agents; animals.

Urquiza-Haas, E. G., & Kotrschal, K. (2015). *The mind behind anthropomorphic thinking: Attribution of mental states to other species*. Animal Behaviour, 109, 167–176. https://doi.org/10.1016/j.anbehav.2015.08.011

IN SHORT WHEN WE INTERACT WITH MACHINES

- **We held intelligent systems on a different moral standard.**
- We **do not** always **understand that we are interacting** with an artefact.
- We **do not** always **understand a system's actions/behaviours.**
- We **do not understand a system's limitations.**

UNIVERSITY OF BATH

WHY IS THIS AN ISSUE?

INCIDENTS

"Alexa, Can I Trust You?"

Hyunji Chung, Michaela Iorga, and Jeffrey Voas, NIST
Sangjin Lee, Korea University

Several recent incidents highlight significant security and privacy risks associated with intelligent virtual assistants (IVAs). Better diagnostic testing of IVA ecosystems can

For ex
6-year-old
love of dc
the famil
prompted
her parer
Kraft. So

RESEARCH ARTICLE

Even good bots fight: The case of Wikipedia

Milena Tsvetkova[1], **Ruth García-Gavilanes**[1], **Luciano Floridi**[1,2], **Taha Yasseri**[1,2]*

1 Oxford Internet Institute, University of Oxford, Oxford, United Kingdom, **2** Alan Turing Institute, London, United Kingdom

* taha.yasseri@oii.ox.ac.uk

Abstract

In recent years, there has been a huge increase in the number of bots online,
Web crawlers for search engines, to chatbots for online customer service, sp
social media, and content-editing bots in online collaboration communities. T

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

SHARE

MOST READ

BBC Sign in News Sport Weather Shop Reel Travel More Search

NEWS

Home Video World UK Business Tech Science Stories Entertainment & Arts Health World News TV More

Technology

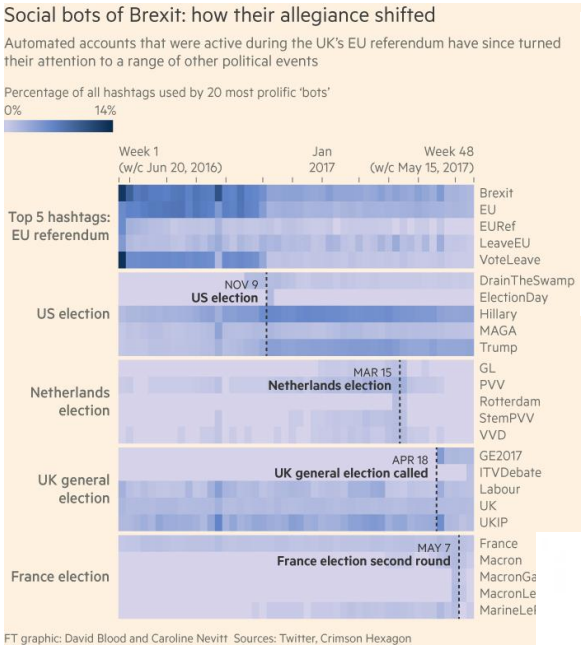
Amazon scrapped 'sexist AI' tool

10 October 2018 Share

Top Stories

US attorney general held in contempt

...AND MORE INCIDENTS



Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum

COMPROP RESEARCH NOTE 2016.1

Bots and Automation over Twitter during the U.S. Election

COMPROP DATA MEMO 2016.4 / 17 NOV 2016

Bence Kollanyi
Corvinus University
kollanyi@gmail.com
@bencekollanyi

Philip N. Howard
Oxford University
philip.howard@oii.ox.ac.uk
@pnhoward

Samuel C. Woolley
University of Washington
samwooll@uw.edu
@samuelwoolley

ABSTRACT

Bots are social media accounts that automate interaction with other users, and political bots have been particularly active on public policy issues, political crises, and elections. We collected data on bot activity using the major hashtags related to the U.S. Presidential Election. We find that that political bot activity reached an all-time high for the... over time, but the ga... the first debate to 5:... the election, most cl... content production a... after Election Day.

DISINFORMATION AND SOCIAL BOT OPERATIONS IN THE RUN UP TO THE 2017 FRENCH PRESIDENTIAL ELECTION

EMILIO FERRARA
UNIVERSITY OF SOUTHERN CALIFORNIA, INFORMATION SCIENCES INSTITUTE

ABSTRACT

Recent accounts from researchers, journalists, as well as federal investigators, reached a unanimous conclusion: social media are systematically exploited to manipulate and alter public opinion. Some disinformation campaigns have been coordinated by means of bots, social media accounts controlled by

Cambridge Analytica

2017 EUROBAROMETER

- **61%** of respondents have a **positive view** of robots
- **84%** of respondents agree that **robots can do jobs** that are too **hard**/**dangerous** for people
- **68%** agree that robots are a **good** thing **for society** because they help people
- **88%** of respondents consider robotics a technology that **requires careful management**
- **72%** of respondents think robots **steal people's jobs**

LIKE THE ELEVATORS

WE NEED TO BUILD TRUST FOR OUR SYSTEMS

- To **perform as we expect them to.**

- The implications from their development and deployment fall within:
  - **E**thical
  - **L**egal
  - **S**ocial
  - **E**conomic
  - **C**ultural

  (**ESLEC**) specifications and values we want to protect.

AI GOVERNANCE

UMEÅ UNIVERSITY


DOWNING
STREET SW1
CITY OF WESTMINSTER
WHITEHALL
SW1
CITY OF WESTMINSTER

EPSRC
PRINCIPLES OF ROBOTICS

1. **Robots are multi-use tools.** Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.
2. **Humans, not robots, are responsible agents.** Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.
3. **Robots are products**. They should be designed using processes which assure their safety and security.
4. **Robots are manufactured artefacts.** They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.
5. **The person with legal responsibility for a robot should be attributed.**

European Union Background on AI

EU STRATEGY ON ARTIFICIAL INTELLIGENCE

published in April 2018

Boost AI uptake	Tackle socio-economic changes	Ensure adequate ethical & legal framework

In this context: appointment of Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) in June 2018

European Commission

Ethics Guidelines for AI – Requirements

Human agency and oversight

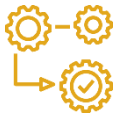
Technical Robustness and safety

Privacy and data governance

Transparency

Diversity, non-discrimination and fairness

Societal & environmental well-being

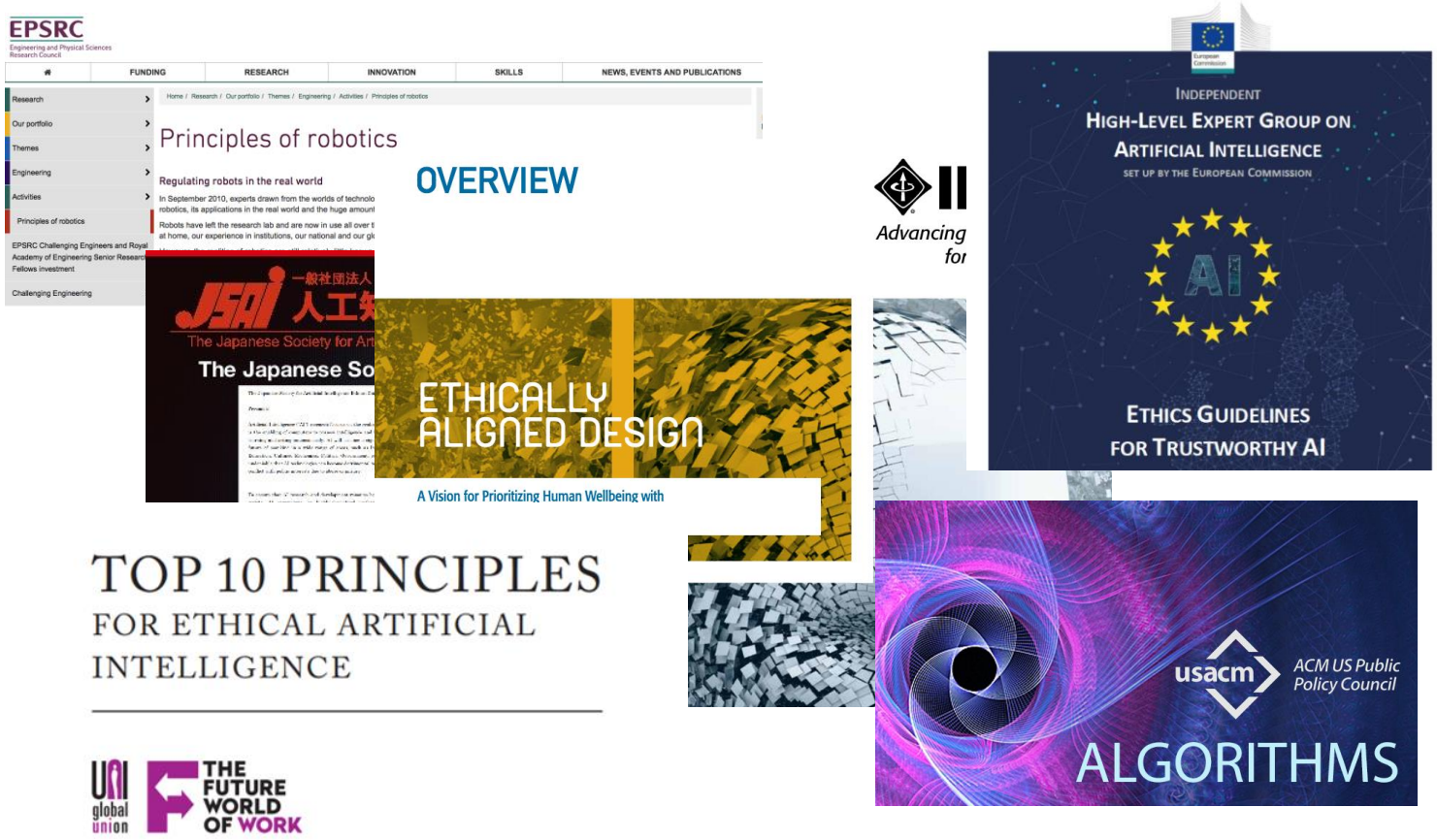


Accountability

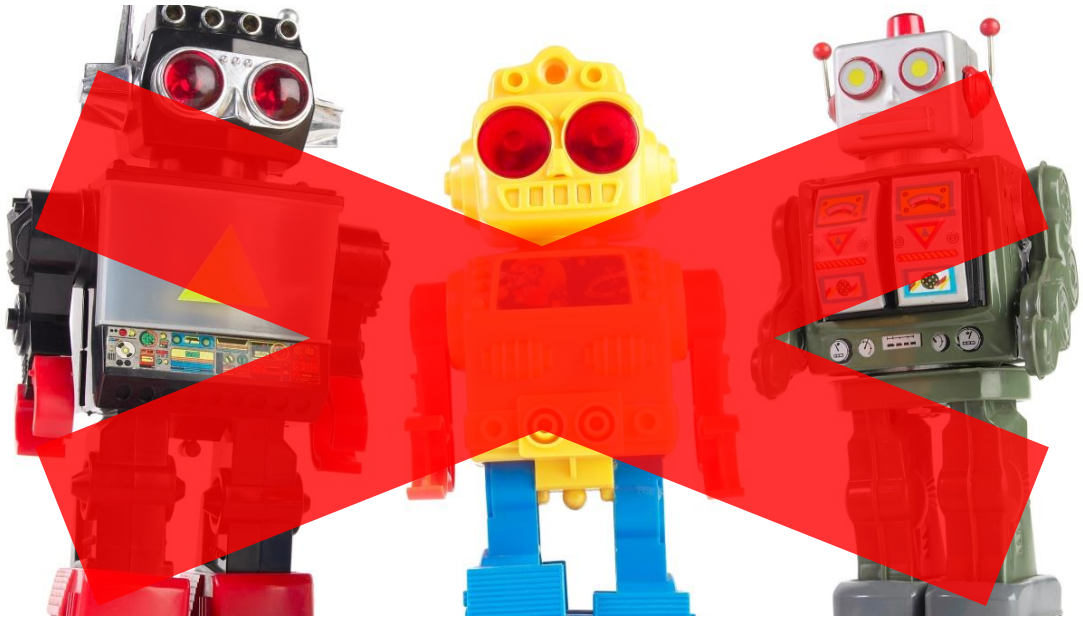
To be continuously implemented & evaluated throughout AI system's life cycle

European Commission

HIGH-LEVEL GUIDELINES



THEY DON'T ARE NOT ADDRESSING THESE:

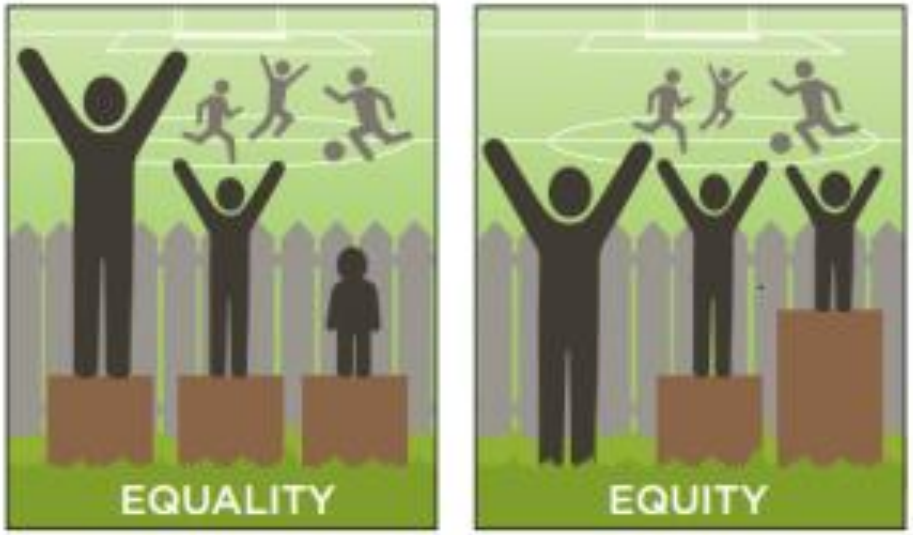


EU HLEG	OECD	IEEE EAD
• Human agency and oversight • **Technical robustness and safety** • Privacy and data governance • **Transparency** • **Diversity**, non-discrimination and fairness • **Societal and environmental well-being** • **Accountability**	• benefit people and the planet • respects the rule of law, human rights, democratic values and **diversity**, • include appropriate safeguards (e.g. human intervention) to ensure a fair and just society. • **transparency** and responsible disclosure • **robust, secure and safe** • Hold organisations and individuals **accountable** for proper functioning of AI	• How can we ensure that A/IS do not infringe **human rights**? • Traditional metrics of prosperity do not take into account the full effect of A/IS technologies on **human well-being**. • How can we assure that designers, manufacturers, owners and operators of A/IS are **responsible and accountable**? • How can we ensure that A/IS are **transparent**? • How can we **extend the benefits** and minimize the risks of AI/AS technology being **misused**?

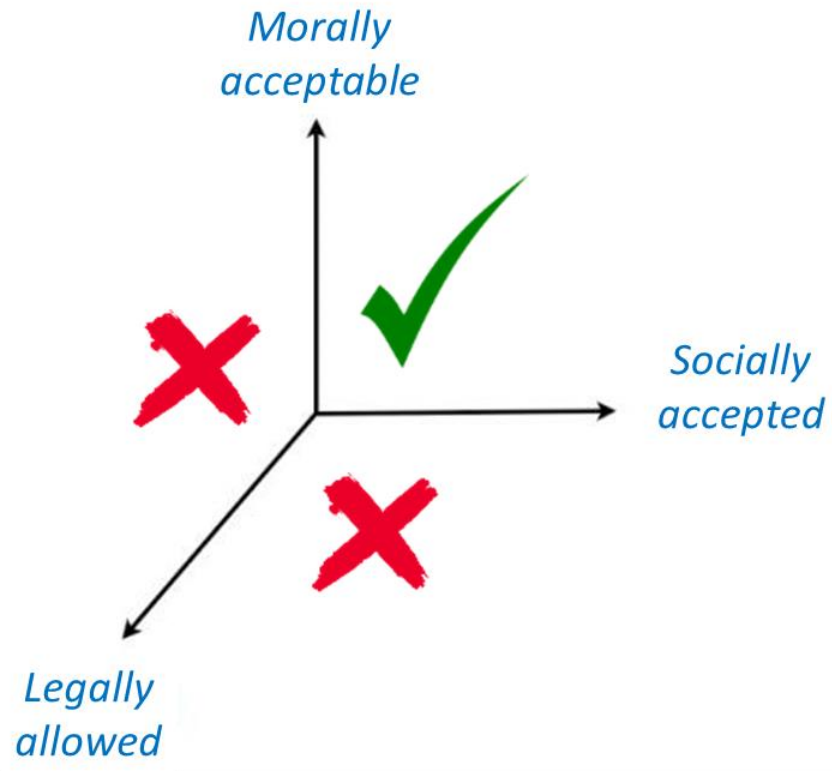
BUT WHAT DO THESE VALUE ACTUALLY MEAN?

HOW DO YOU INTERPRET THEM?

- Values have **different interpretations** in different contexts and cultures.

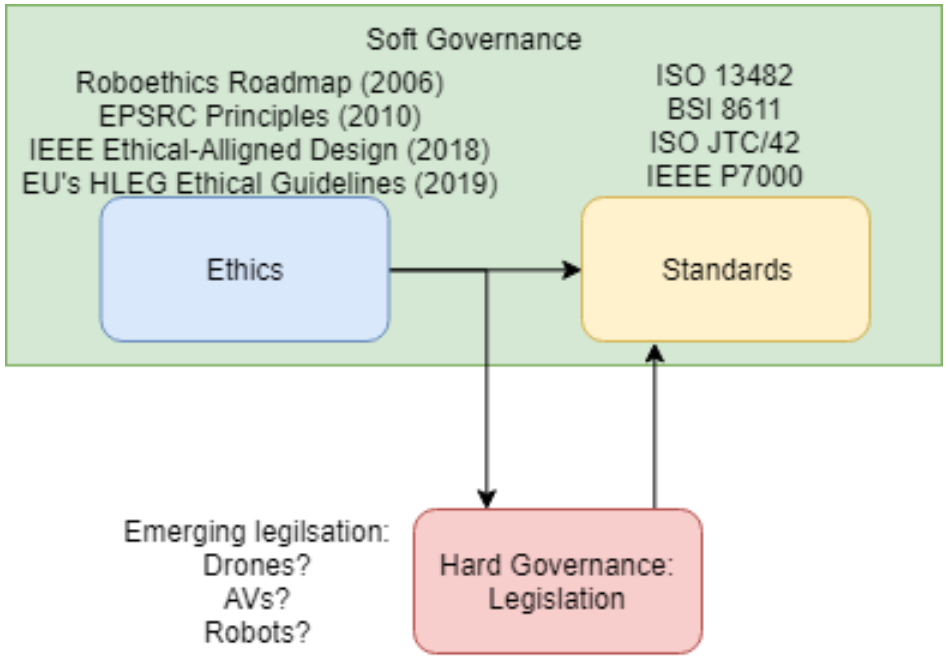


THIS INTERPRETATION NEEDS TO CONSIDER



THE NEED TO AUDIT

- Only when these interpretation are clear, we can talk about actual Governance.



PROMOTING GOVERNANCE

- For effective governance, we **need to be able to audit our systems** to:
  - find out what went wrong and why,
  - debug our systems;
  - Check compliance of a system adheres to our values.

- Sensible implementation of **transparency** can help us achieve that.

Theodorou A. (2019). *AI Governance Through a Transparency Lens. PhD Thesis*. University of Bath, UK
Bryson J.J., **Theodorou A.** (2019). *How Society Can Maintain Human-Centric Artificial Intelligence*. Toivonen-Noroand M and Saari E eds. *Human-Centered Digitalization and Services*. Springer, Berlin.

ACCOUNTABILITY
RESPONSIBILITY
TRANSPARENCY

UMEÅ UNIVERSITY

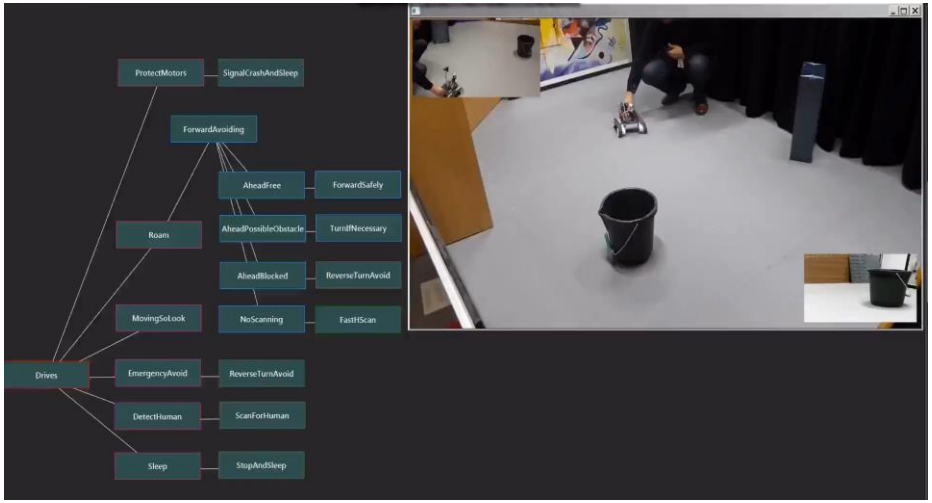
WHAT IS "TRANSPARENCY"?

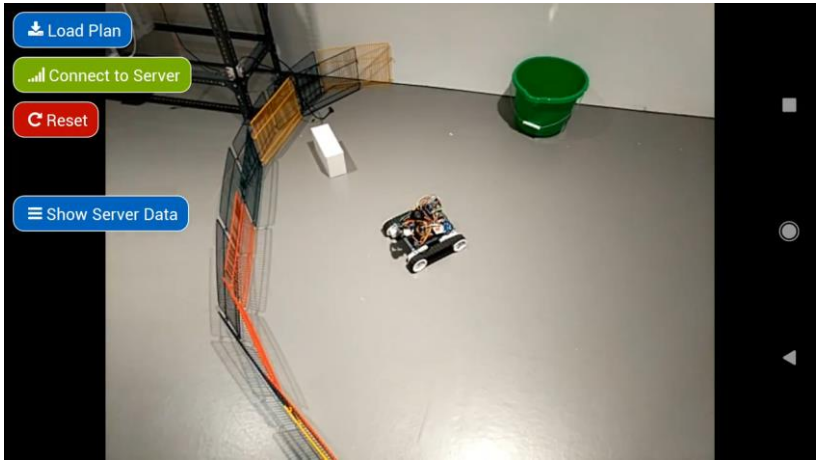
- The decision-making mechanism should be exposed.

- Available on-demand, at any point of time, accurate interpretations of:
  - goals,
  - process towards goals,
  - sensory inputs, and
  - unexpected behaviour.

Theodorou A., Wortham R.H., and Bryson J. *Designing transparency for real time inspection of autonomous robots*. Connection Science, Vol. 29, Issue 3

UNIVERSITY OF BATH





Theodorou A. (2017), *ABOD3: A Graphical Visualization and Real-Time Debugging Tool for BOD Agent*. CEUR Workshop Proceedings, 1855, pp. 25-30.

Rotsidis A., **Theodorou A.,** and Wortham R.H., 2019. *Robots That Make Sense: Transparent Intelligence Through Augmented Reality,* 1st International Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies. Los Angeles, CA USA.

OPAQUE VS TRANSPARENT

Results (N=40)	Group One (w/o ABOD3)	Group Two (ABOD3)
Robot is thinking	**0.36 (SD 0.48)**	**0.65 (SD 0.48)**
Robot is intelligent	2.64 (SD 0.88)	2.74 (SD 1.07)
Understanding Objective	0.68 (SD 0.47)	0.74 (SD 0.44)
Mental Model Accuracy	**1.86 (SD 1.42)**	**3.39 (SD 2.08)**

Wortham, R.H., **Theodorou, A.** and Bryson J.J., (2016).*What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems*, IJCAI-2016 Ethics for Artificial Intelligence Workshop, New York USA

OPAQUE VS TRANSPARENT

Results (N=55)	Group One (w/o ABOD3)	Group Two (ABOD3)
Robot is thinking	0.46 (SD 0.50)	0.56 (SD 0.50)
Robot is intelligent	2.96 (SD 0.18)	3.15 (SD 1.18)
Understanding Objective	0.50 (SD 0.50)	0.89 (SD 0.31)
Mental Model Accuracy	**1.89 (SD 1.42)**	**3.52 (SD 2.10)**

Wortham, R.H., **Theodorou, A.** and Bryson J.J., (2017). *Improving Robot Transparency: Real-Time Visualisation of Robot AI Substantially Improves Understanding in Naive Observers*, IEEE RO-MAN 2017, Lisbon, Portugal

ACCURATE MENTAL MODELS

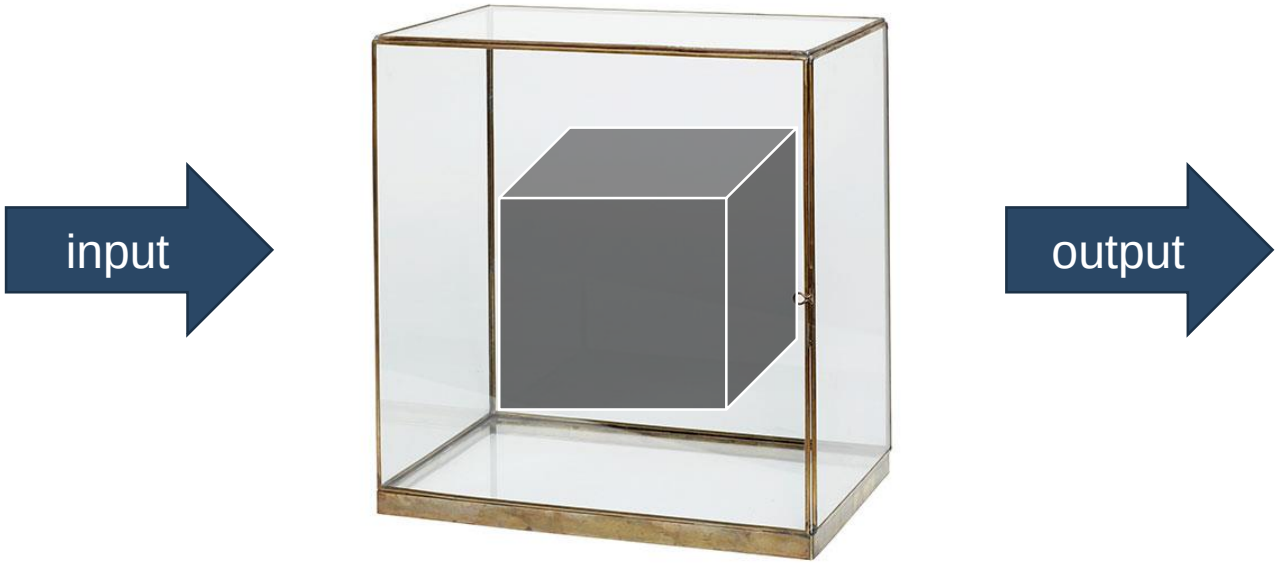
- **Misunderstanding** leads to anxiety, mistrust, fear and **misuse/Disuse**
- **User self doubt** – "What is going on here? Is the robot supposed to do this or **did *I* do something wrong**?" *
- With poor Transparency, robots that **can mislead** us. *
- With good Transparency, we can **calibrate trust** (choose to **trust** or **lose confidence**)

* Taemie Kim and Pamela Hinds (2006). *Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction*, Proceedings - IEEE International Workshop on Robot and Human Interactive Communication, 80–85, (2006).
*2 P. a. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman. *A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction*, Human Factors: The Journal of the Human Factors and Ergonomics Society, 53(5), 517–527, (2011).

KEEPING THE BLACK BOX

- **Sometimes black boxes are inevitable.**
- Some of the best performing methods for pattern recognition, e.g. deep learning, are black boxes right now.
- Yet, we still need to audit our systems.
- **Traceability of all decisions is necessary; that starts with your policy and goes to usage.**

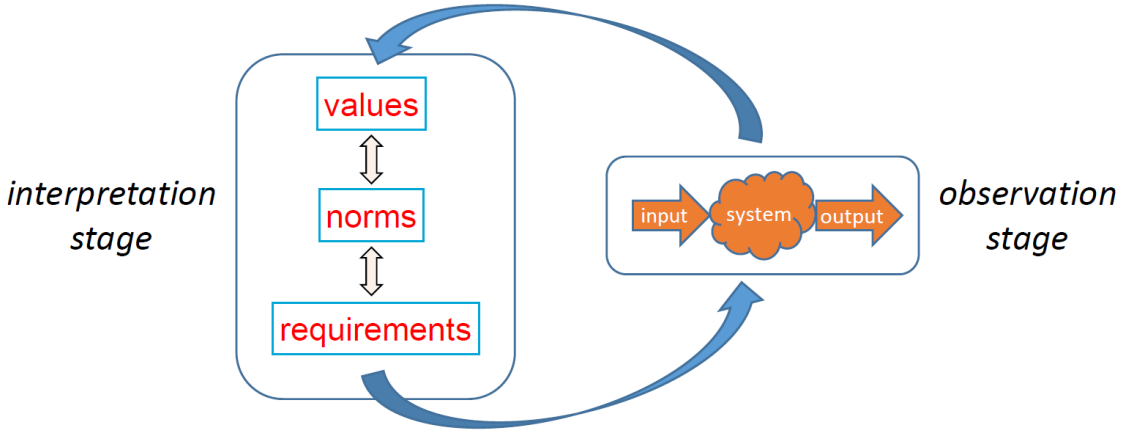
GOVERNANCE BY ~~BLACK~~ GLASS BOX



Aler Tubella A., **Theodorou A.**, Dignum F., Dignum V. (2019). *Governance by Glass-box: implementing transparent moral bounds for AI behaviour*. International Joint Conference on Artificial Intelligence (IJCAI) 2019. Macao, China.

TWO-STAGES SYSTEM

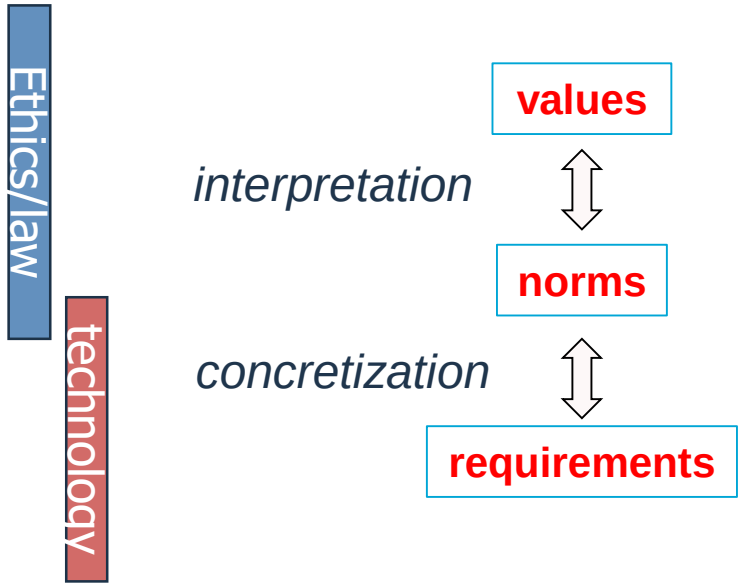
- Checks whether a system adheres to ESLEC values.



Aler Tubella A., **Theodorou A.**, Dignum F., Dignum V. (2019). *Governance by Glass-box: implementing transparent moral bounds for AI behaviour*. International Joint Conference on Artificial Intelligence (IJCAI) 2019. Macao, China.

INTERPRETATION STAGE

- Structured and explicit process of translating translate **abstract values** into **concrete norms** and **requirements**.
- We aim to not only **describe** the norms themselves, but also **the exact connection** between abstract and concrete concepts **in each context**.
- Fulfilling the norm will be considered as adhering to the value.

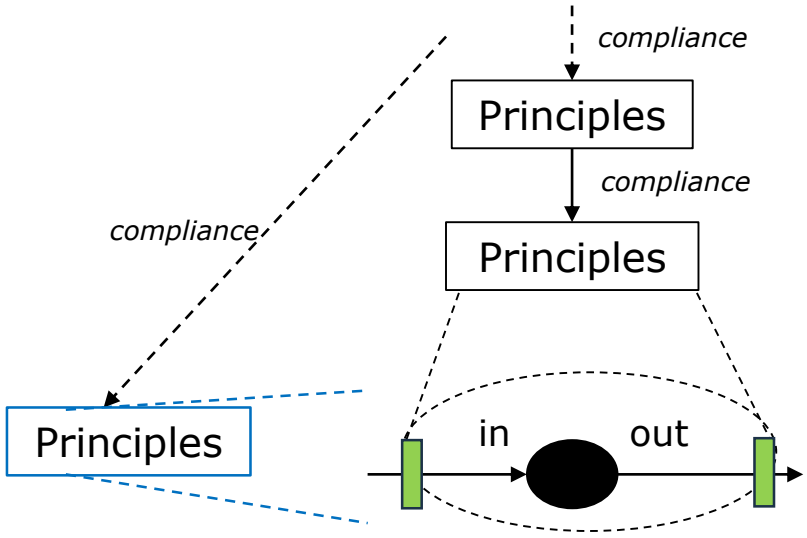


OBSERVATION STAGE

- **Continuously monitoring** the system by using our interpretation-stage requirements to **define and perform tests**.

- Explicitly **showcase** which **values** are **being met**, in which context and how.

- Allows us to **enforce our values**: accept or not a system's decision.

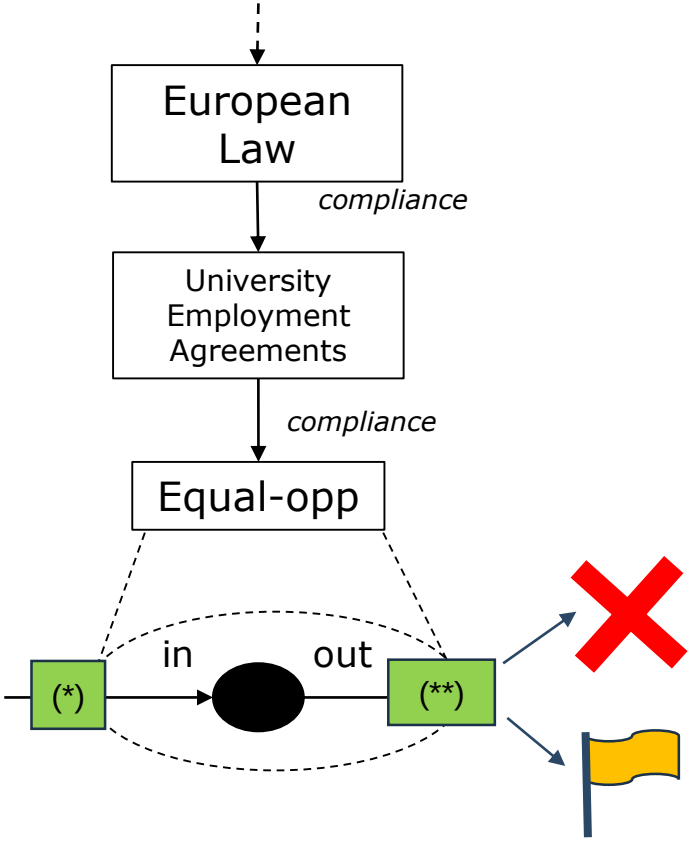


EXAMPLE: RECRUITMENT SYSTEMS

- Value: Fairness
- Norm: Equal opportunity
- Implementation:
  - Input

    (*) Gender of candidate not in input
  - Output evaluation

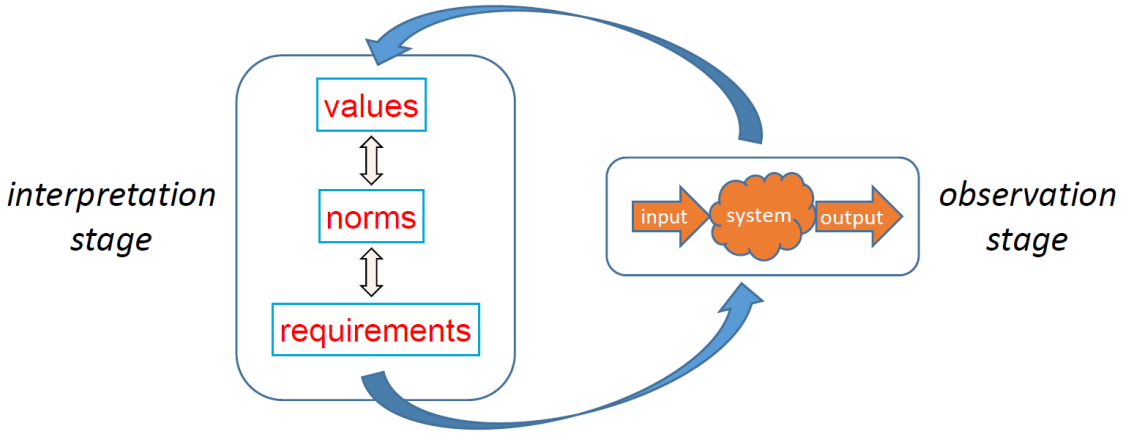
    (**) P(job | female) = P(job | male) every N decisions
- Governance
  - Cut-off
  - Flag-out



TWO-STAGES SYSTEM

- The two stages inform each other.

- Results from the observation may tune the interpretations --- and the system itself.



FORMALISING THE GLASS BOX

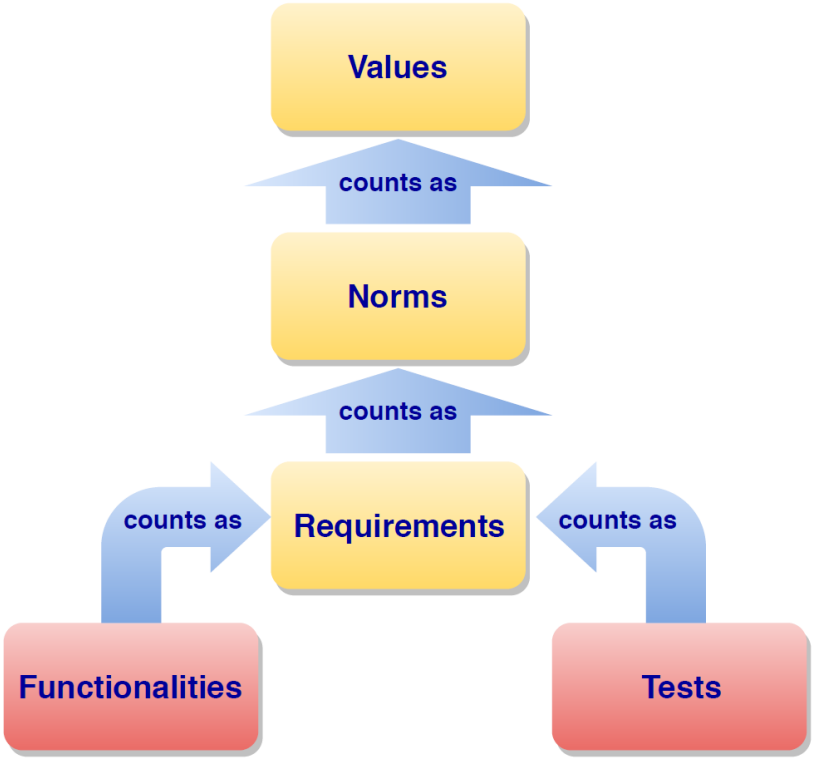
Andreas Theodorou | t: @recklesscoding

CHALLENGES & AIMS

- **Domain-agnostic**, to allow for adaptation to any application.
- **Context-aware**, to explicitly describe in which context a functionality relates to a value.
- **Implementable**, able to be encoded in a programming language.
- **Computationally tractable**, to allow for verification and monitoring in reasonable time.

FORMALISING THE GLASS BOX

- A multi-modal logic with *counts-as operator* is enough to encode a Glass Box.
- We encode statements of the form: "***A* counts-as *B* in context *C***".
- It allows for verification in reasonable time.



Aler Tubella A., Dignum V. (2019). *The Glass Box Approach: Verifying Contextual Adherence to Values*. Workshop in AI Safety 2019

TRANSPARENCY IS NOT EVERYTHING

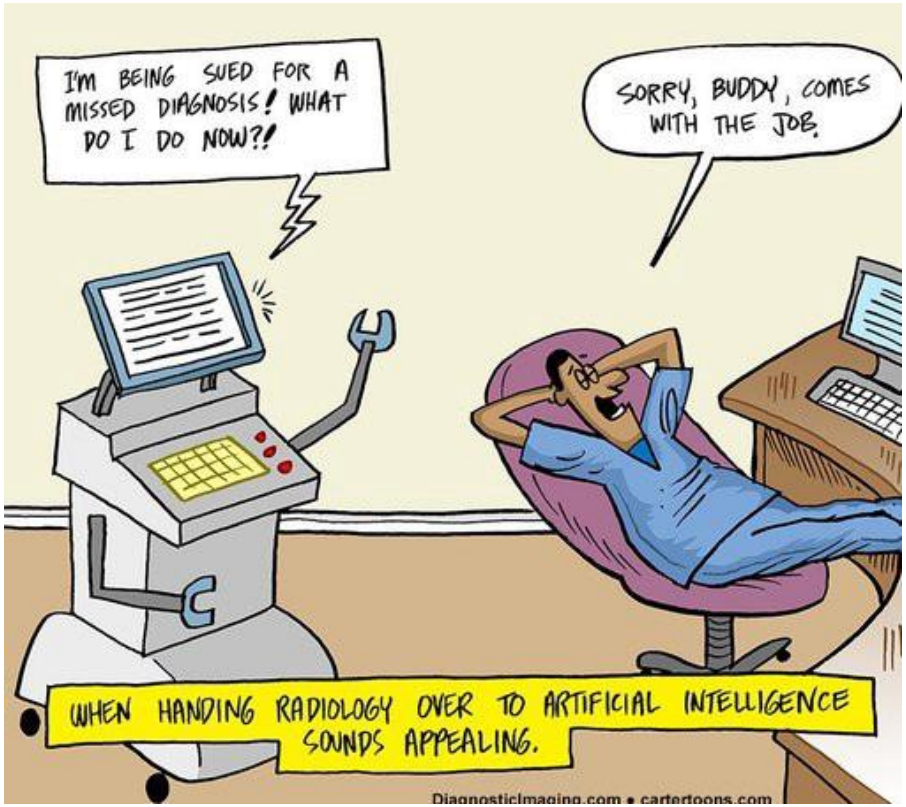
- Transparency **is not** the **end goal**.
- Transparency is just a "tool" to help us find out *what* went wrong (Theodorou, 2017).
- The end goal is **responsibility and accountability** (Bryson, 2019).

Theodorou A., Wortham R.H., and Bryson J. *Designing transparency for real time inspection of autonomous robots*. Connection Science, Vol. 29, Issue 3

Bryson J.J., **Theodorou A.** (2019). *How Society Can Maintain Human-Centric Artificial Intelligence*. Toivonen-Noroand M and Saari E eds. *Human-Centered Digitalization and Services*. Springer, Berlin.

RESPONSIBILITY

- **Responsibility** refers to **the role of people themselves** and to the capability of AI systems to answer for one's decision and identify errors or unexpected results.
- There is a "**chain of responsibility**".
- We are *moral agents*, never the machines.



ACCOUNTABILITY

- When things go wrong, we may held individuals accountable.
- Accountability is not just about “punishing”, it is also about addressing issues (sometimes readdressing).
- The “threat” of legal liability motivates organisations (and individuals) to demonstrate their *due diligence*.
- **Your policy, your decisions, your system form your due diligence.**

DOES RESPONSIBLE AI SOUND EASY-ISH?

It is not.

IT IS A LONG & HARD PROCESS

"We've got to draw the line on unethical behavior. But draw it in pencil."

69. Moral functionalism (also **instrumentalism**)

The view that ethics should merely be a useful instrument for other purposes. A risk is that ethics is not seen as a value in and of itself.

IT INVOLVES

- Ethics **in** Design: **Development** is **influenced** by **ESLEC** issues.

- Ethics **by** Design: **Integration** of **ethical abilities** as part of the **behaviour** of artificial intelligent **systems**.

- Ethics **for** Design: Codes of conduct, standards, and certification processes that **ensure** the **integrity** of **developers** and **users**.

Dignum, V (2018). *Ethics in Artificial Intelligence: Introduction to the special issue*. Ethics and Information Technology, 20(1):1–3, 3 2018.

ETHICS IS NOT AN AFTERTHOUGHT

Not a checklist based on some high-level guidelines to tick once and forget.

CONTEXT MATTERS

Stakeholders, projects, societies that will be deployed to, etc should be taken into consideration through the process.

How YOU interpret any ESLEC values needs to be clear.

OH! AND AVOID OVERSTATEMENTS.

You can’t have an “unbiased” data-driven system. It simply wouldn’t work.

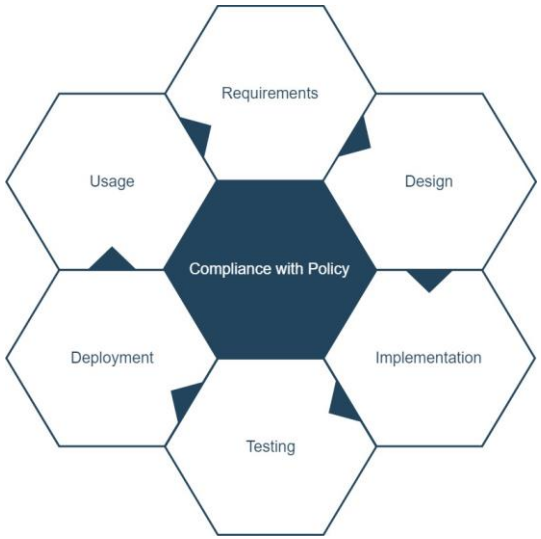
OH! WE CAN ALSO HELP!

AI4EU

- Europe's prominent *on-demand AI platform*.
- Aims to **help Small-Medium Enterprises access tools and expertise** across the Union.
- The catch is that it **promotes** the development ***Responsible AI***.

AI4EU'S RESPONSIBLE AI METHODOLOGY

- Policy becomes the centre of a system's life cycle.
- Promote **compliance** with **both *legal*** and ***ethical*** policy.
- Help make responsible AI **part** of the organisation's **culture**.



Umeå University Responsible AI Group

Virginia Dignum
Professor Social and Ethical AI
@vdignum
virginia@cs.umu.se

Andrea Aler Tubella
Postdoctoral Researcher
andrea.aler@umu.se

Andreas Theodorou
Postdoctoral Researcher
@recklesscoding
andreas.theodorou@umu.se

Zahoor Ul Islam
PhD Student
zahoor.ul.islam@umu.se

Julian Mendez
PhD Student
julian.mendez@umu.se

AI4EU

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Co-authors @ Umeå University

Frank Dignum
Professor Social Aware AI
frank.dignum@umu.se

Juan Carlos Nieves
Assoc. Professor
juan.carlos.nieves@umu.se

International Co-authors

Joanna J Bryson
Professor of Technology & Ethics
Hertie School of Governance Berlin
@j2bryson

Robert H. Wortham
Teaching Fellow – University of Bath
@rwortham

Read our group's research:

Holly Wilson
PhD Student – University of Bath

UMEÅ UNIVERSITY

Q & A

Andreas Theodorou

andreas.theodorou@umu.se | @recklesscoding

Read our group's research: