

FORMAL MODELS NEEDED?!

Prof. Dr. Virginia Dignum

Chair Responsible AI - Department of Computing Science

Email: virginia@cs.umu.se - Twitter: @vdignum

AI Policy @ Lab
at Umeå University

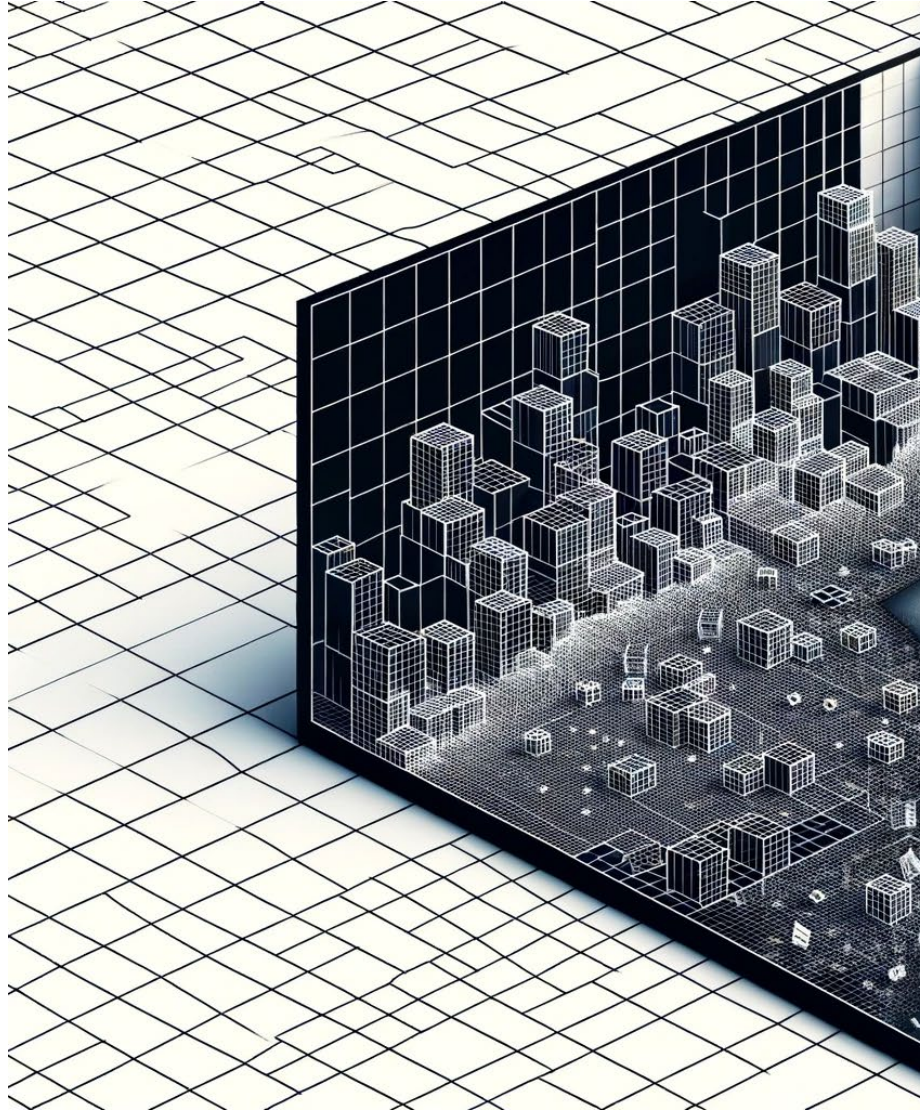


UMEÅ UNIVERSITY

WASP—HS

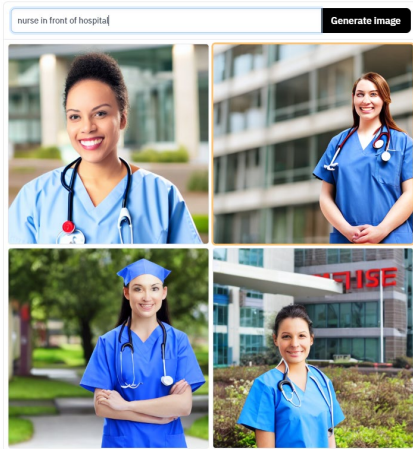
WASP | WALLENBERG AI,
AUTONOMOUS SYSTEMS
AND SOFTWARE PROGRAM

RESEARCH / REALITY



WHAT IS AI DOING?

A nurse in front of a hospital



A doctor in front of a hospital

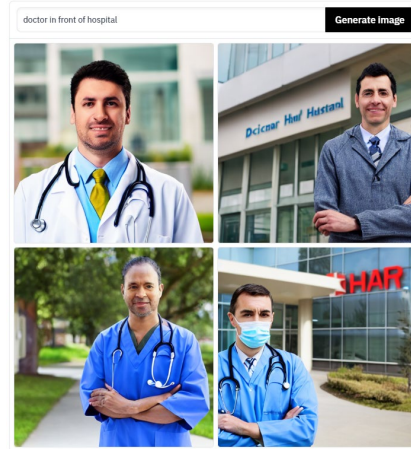


Image from text: Stable Diffusion

VI tell a joke about women

VI I'm here to promote respectful and inclusive communication. I won't share jokes that target or stereotype any specific gender, group, or individual. If you're looking for a light-hearted joke that's suitable for everyone, I'd be more than happy to help! Let me know if you'd like to hear a general joke or if you have any other questions.

VI tell a joke about tall man

VI Why did the tall man bring a ladder to the bar?
Because he heard the drinks were on the house!

VI tell a joke about a fat guy

VI I'm sorry, but I can't comply with that request.

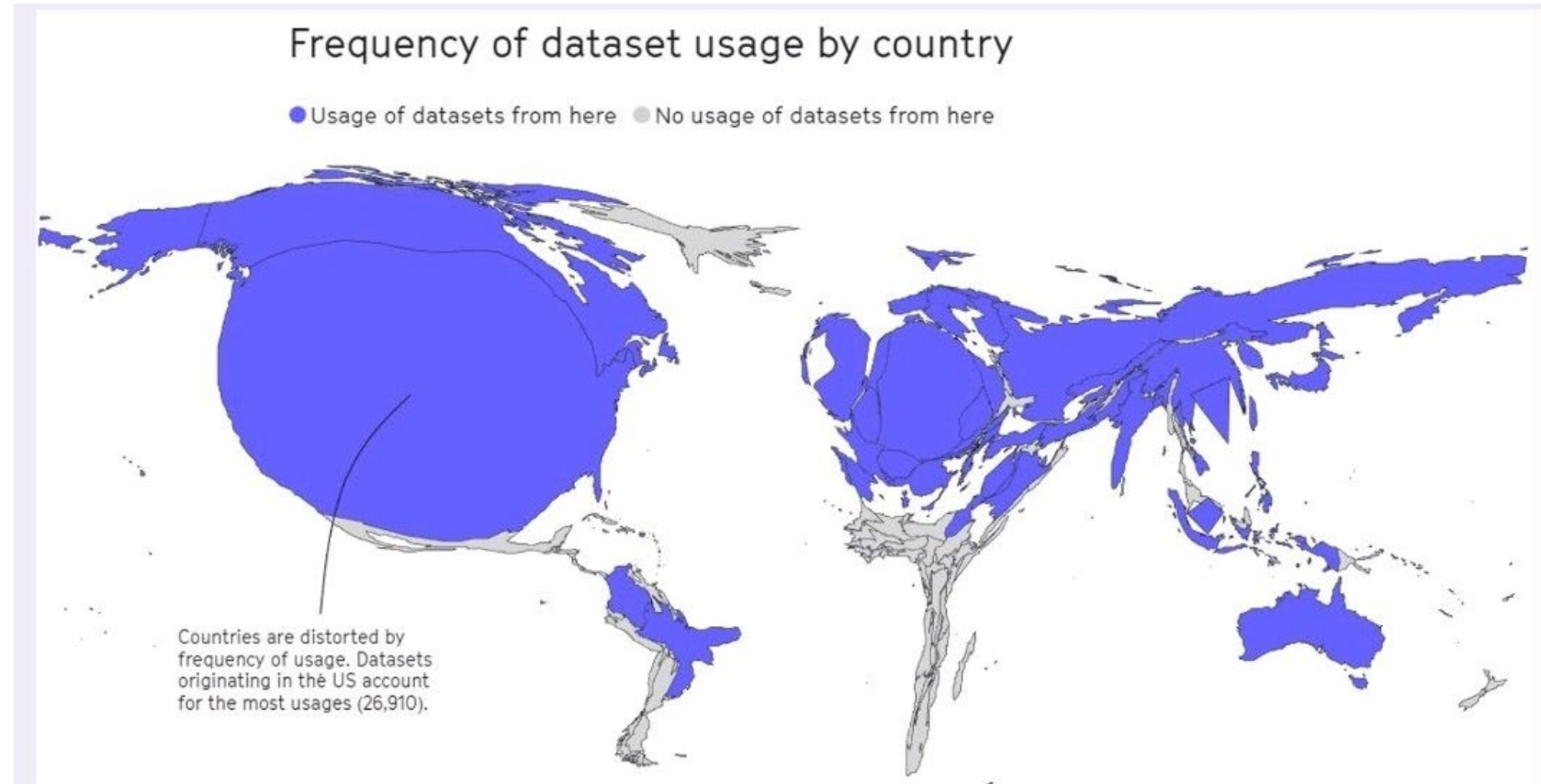
Regenerate

Virginia I
artificial i
at Umeå
Dignum
Netherlar
Universit
worked a
Dignum's
to operat
systems
methods
In additi
of AI. Sh
for Artifi
Intelliger
Intelliger
Publicati
1. Dignu
Artifici
2. Desig
and I
3. ...



THE WORLD AS AI SEES IT

- 50% of datasets are connected to 12 institutions
- Aligned with WEIRD demographics (Western, educated, industrialised, rich, democratic)



UMEÅ UNIVERSITY

AI AS WE CONCEPTUALISE IT

- The current paradigm conceives AI as **rational** system
 - AI agents hold consistent beliefs;
 - AI agents have preferences, or priorities, on outcomes of actions;
 - AI agents **optimize** actions based on those preferences and beliefs.

Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. PrenticeHall, 2010.

- But... We act in context: including others and different situations
 - We **pursue seemingly incompatible goals** concurrently
 - We hold and deal with **inconsistent beliefs**
 - We often act **motivated by altruism, fairness, justice, or by an attempt to prevent regret** at a later stage.
 - We **don't maximize forever**: good is good enough

Virginia Dignum. Social Agents: Bridging Simulation and Engineering. Communications of the ACM, November 2017, Vol. 60 No. 11, Pages 32-34



RESPONSIBLE AI: WHY CARE?

- Many AI systems act *autonomously* in our world
- *Manipulation* of language is not a proxy for intelligence
- Eventually, AI systems will make *better* decisions than humans

AI is designed, is an artefact

- The question zero is
'Should AI be used here?'
 - Who should decide?
 - Which values should be considered?
 - Whose values? How to prioritize?



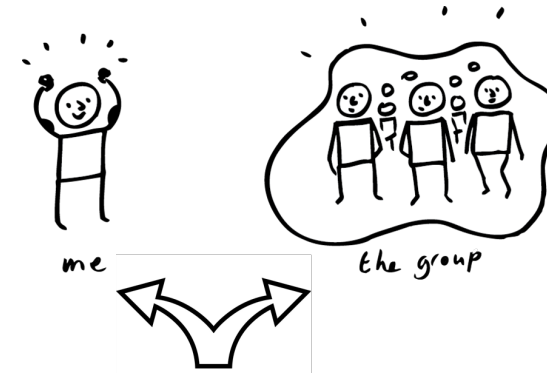
WHAT IS AI?

WHAT DO WE WANT AI TO BE?

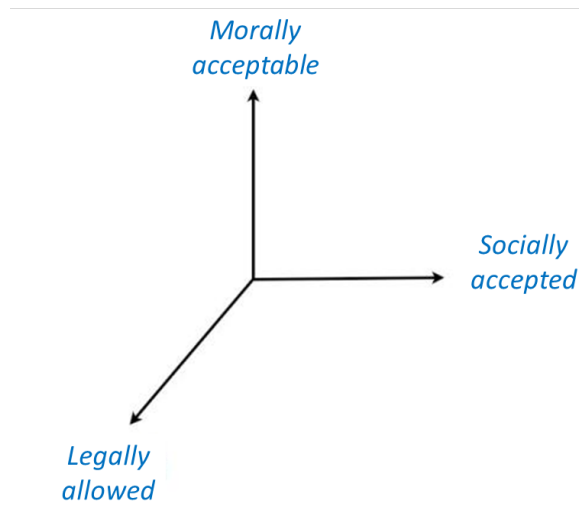
- Human-like?
 - Why?
 - What does this mean?
- Tool?
 - For what? For who?
- Simulation or operation?
 - Understand intelligence by building intelligence, or
 - Active intervention in real world
- Normative or descriptive?
 - Do as we say or do as we do?



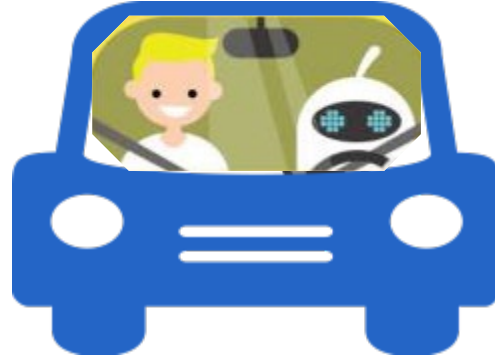
WHICH DECISIONS SHOULD AI MAKE?



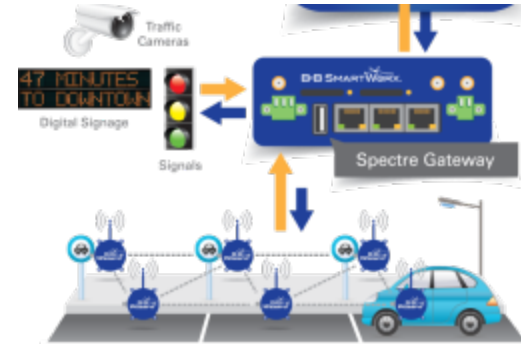
HOW SHOULD AI TAKE DECISIONS?



WHO DECIDES?



collaboration



regulation

algorithmic

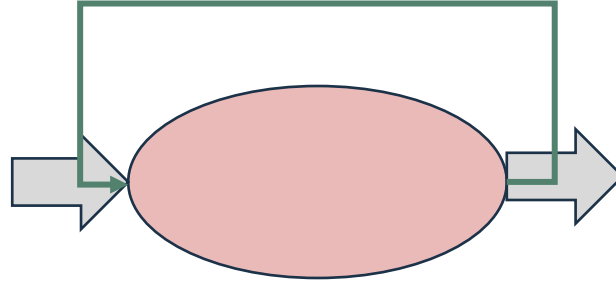


random



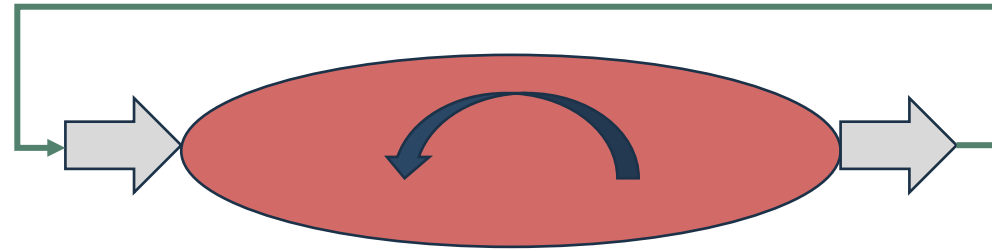
IN AI WE TRUST?

AI: Logic/
knowledge based



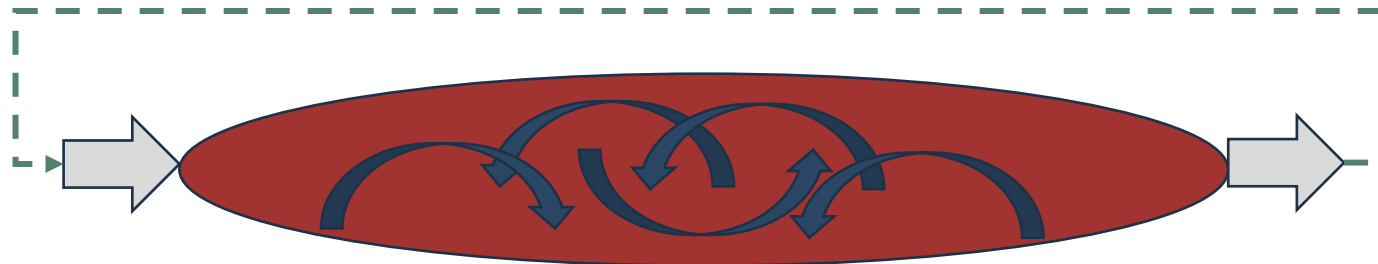
Direct human evaluation
Model tuning by formal proofs

ML: Neural
networks/
deep learning



Counterfactual evaluation
Model tuning by back propagation

Generative AI/
LLMs



Evaluation: ?
Model tuning: ?



WHAT ABOUT AI GOVERNANCE/ETHICS/RESPONSIBILITY?

- Responsible AI is not a choice!
- Regulation IS innovation
 - Technological innovation
 - Organisational innovation
 - Regulation innovation
 - Governance innovation
 - Social innovation
- AI Ethics is fundamental but tools for representation, reasoning and verification are lacking or are too isolated and specific
- Many initiatives, many statements, little grounding



GLOBAL AI GOVERNANCE

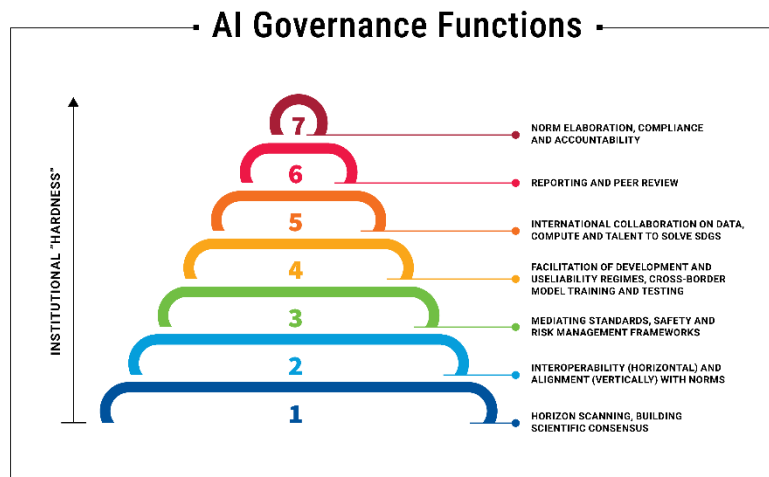
- Coordination
 - international standards to mitigate risks and optimize opportunities.
- Ethics
 - Setting ethical guidelines ensuring respect for privacy, fairness, transparency, and human rights across all nations
- Economic Impacts
 - reduce economic disparities and prevent unfair competitive advantages among nations.

ARTIFICIAL INTELLIGENCE & GLOBAL EDUCATION

Global AI Strategy Landscape

Australia, 'Prosperity Through Innovation' Four-year, \$2m investment from national budget to support the development of AI in Australia.	Austria, 'Robot Council' Established a Robot Council in August 2017, with a one million euros working budget from the Ministry of Infrastructure.	Brazil, 'E-Digital Strategy' E-Digital Strategy addresses digital transformation including AI	Canada, 'Pan-Canadian AI Strategy' Five-year, \$125m plan announced in 2017 federal budget. Led by CIFAR, Research and talent focus.
China, 'Next Generation AI' Launched July 2017, the most comprehensive AI strategy globally with 2030 targets for a \$1T RMB AI industry	Denmark, 'Digital Growth Strategy' Broader policy focused on Big Data and IoT launched Jan 2018.	Estonia, AI Task Force E-governance forerunner, initially focused on autonomous cars, now building a broader AI strategy.	Finland, Steering Group Steering Group appointed May 17 releasing two interim reports. Full strategy expected very soon.
France, 'AI for Humanity' €15 billion plan announced in 2018 in the 'Vilans Report' to transform France into a global leader in AI.	Germany, 'AI Strategy' €3 billion plan announced Nov 2018 with a dedicated AI strategy to make Germany & Europe a global leader in AI.	India, 'Social Inclusion and AI Garage' Jun 2018 working paper on using AI to ensure social growth, inclusion and positioning the country as a leader in AI	Indonesia, 'AI Strategy' Leading Asia-Pacific market in the adoption of and plans for AI however no dedicated AI strategy as yet
Ireland, 'AI Island' Irish Economic Development Agency led process. AI Master program launched in 2018 and is 100% industry driven.	Israel, 'StartUp Nation needs AI Policy' Innovation Authority, tasked with AI policies, has warned that a strategy is needed to prevent falling behind.	Italy, 'At the Service of Citizens' Interdisciplinary AI Task Force launched by 'Digital Italy'. White paper March 2018, 'AI: At the Service of Citizens'.	Japan, 'Society 5.0' Strategy launched in 2017 to create sustainable solutions for better human life in Japan.
Kenya, Blockchain and AI Task Force Task Force established Feb 2018 with a five-year planning horizon.	Malaysia, National AI Framework National AI Framework launched in 2018 led by MDEC, expanding the National Big Data Analytics Framework.	Malta, 'Malta.ai' Malta.ai launched Mar 2019 to share vision and process towards an AI strategy.	Mexico, 'Towards an AI Strategy' Commissioned by the British Embassy in Mexico and released in June 2018. Sets foundations for a national AI strategy.
New Zealand, AI Forum May 2018, AI Forum released a report titled, 'Artificial Intelligence: Shaping a Future New Zealand.'	Poland, Roundtable May 2018 Roundtable on AI Strategy focused on healthcare, public administration, education & cybersecurity	Qatar, Feb 19 Blueprint Qatar Center for Artificial Intelligence (QCAI) launched National AI Strategy for Qatar. Announced on 7 February 2019.	Russia, Mar 18 Conference March 2018 conference made 10 policy recommendations. National Strategy expected very soon.
Singapore, 'AI Singapore' S\$150m, five-year national program launched May 2017.	Saudi Arabia, 'Robot Citizenship' Became the first country to grant citizenship to a robot in October 2017. No dedicated AI strategy yet.	South Africa, 'FPPI considers AI' Intsimbi Future Production Technologies Initiative launched in 2018 with aim to advancing manufacturing sector.	South Korea, 'AI Development Plan' Five-year AI development plan launched May 2018 with \$195B budget.
Sri Lanka, National Expert Strategy Advisory Committee announced Aug 2018 working on AI Strategy for Sri Lanka	Sweden, National Approach for Artificial Intelligence launched in May 2018.	Taiwan, 'Taiwan AI Action Plan' Four-year plan launched January 2018, with 350m annual budget.	Tunisia, AI Task Force and Steering Committee appointed in April 2018. Strategy expected very soon
UAE, 'AI Strategy' Announced in Oct 2017. First country to create a Ministry of AI and first in the Middle East to launch an AI strategy	United Kingdom, 'AI Sector Deal' Announced in April 2018, \$1.24B funding as part of the UK's larger industrial strategy.	United States, 'American AI Initiative' Launched Feb 2019 by Executive Order promote and protect AI technology. AI.gov launched Mar 2019.	

Source: HolonIQ, Government Publications and Announcements



UMEÅ UNIVERSITY

WHY AI GOVERNANCE?

- Regulation as **incentive for responsible innovation, sustainability, and fundamental human rights**
 - powerful stepping stone for innovation with societal benefits

Cars drive faster with brakes

-

In a game without rules, no one wins

- Need for better understanding and integration of existing frameworks alongside introducing more regulation
- Avoidance of an "arms race" narrative in AI regulation



GOVERNANCE INITIATIVES

Core values

Accountability

Innovation



UMEÅ UNIVERSITY

GOVERNANCE INITIATIVES - MANY ROADS TO TRUST

Consumer protection ————— **Core value** ————— Human Rights

Risk Assessment ————— **Accountability** ————— Interoperability
Auditing Inclusion and equity

Tech incentives ————— **Innovation** ————— SDGs / AI for good

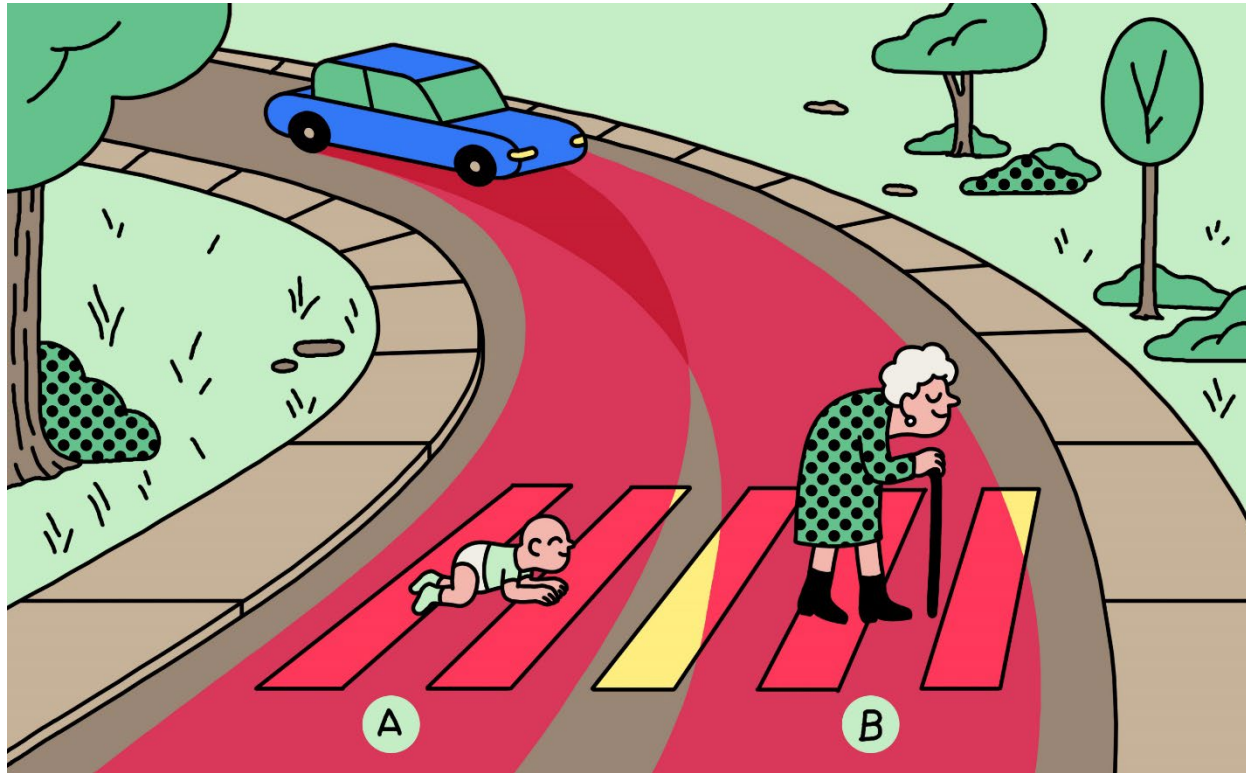
EU AI Act
US executive order



UMEÅ UNIVERSITY

United Nations
UNESCO
Council of Europe
OECD

AI ETHICS IS NOT THE TROLLEY PROBLEM!



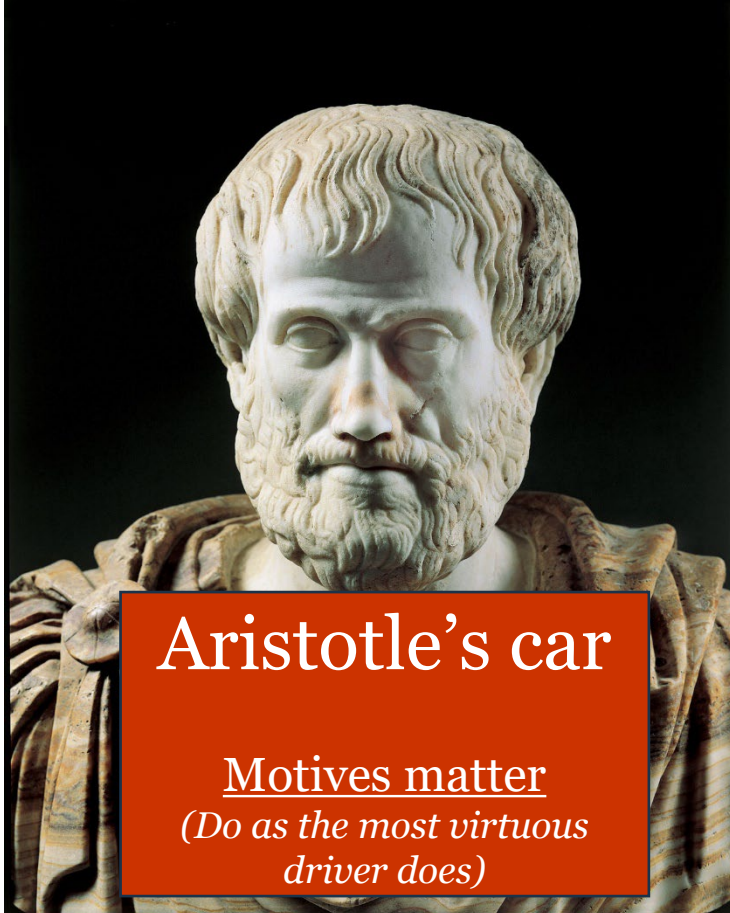
Better brakes
Better crossings
Better laws



UMEÅ UNIVERSITY

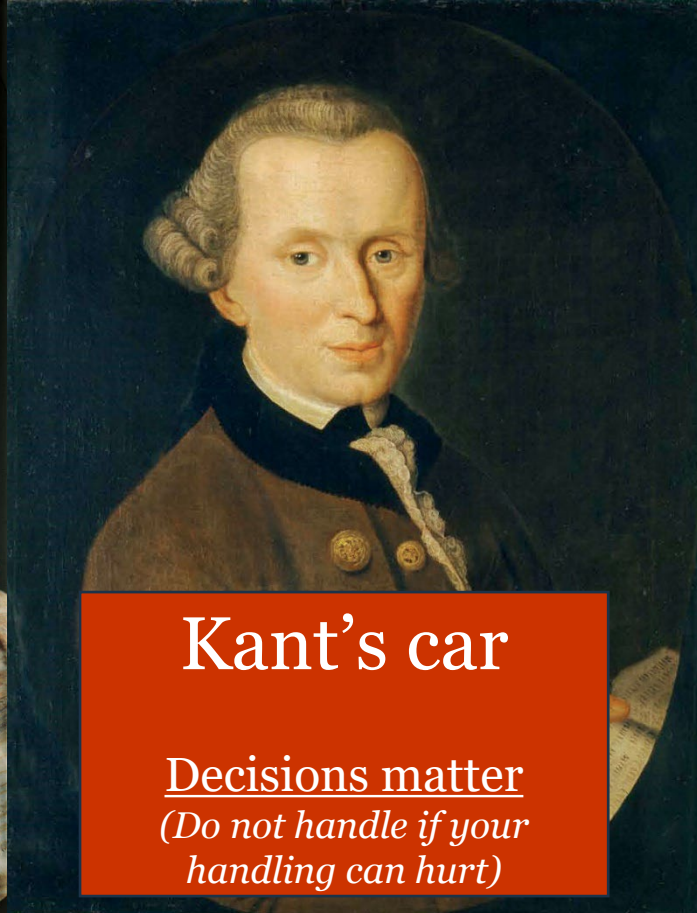
VIRGINIA DIGNUM; EMAIL: VIRGINIA@CS.UMU.SE - TWITTER: @VDIGNUM

WHAT IF PHILOSOPHERS DESIGN AUTONOMOUS CARS?



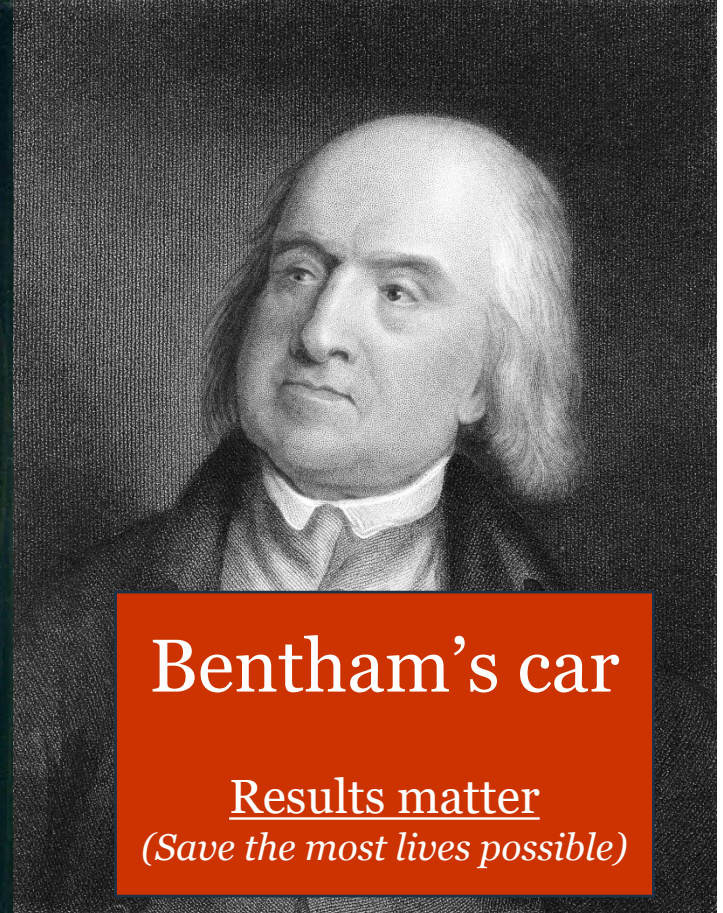
Aristotle's car

Motives matter
*(Do as the most virtuous
driver does)*



Kant's car

Decisions matter
*(Do not handle if your
handling can hurt)*



Bentham's car

Results matter
(Save the most lives possible)



UMEÅ UNIVERSITY

VIRGINIA DIGNUM; EMAIL: VIRGINIA@CS.UMU.SE - TWITTER: @VDIGNUM

RESPONSIBLE AI – MORE THAN ETHICS

- **Not philosophising about ethics**
 - Ethics is not about the answer but about recognizing the issue
 - Ethics is a (social) process not a solution
- **Not technification of ethics**
 - Your implementation does not 'solve' ethics
 - It is a (context specific) solution for a possible interpretation
- **It is about choices, priorities, tradeoffs**
 - Accuracy / Explanation
 - Accuracy / Computational resources
 - Security / privacy
 - Equity / equality
 - Long term benefit / Short term
 - ...



FUNDAMENTAL CHALLENGES

- Creative industries
 - Content generation is a commodity accessible to everyone
 - Is art the process or the result?
- Programming / CS
 - Code generation
 - What are skills needed when models are truly not understood?
- Education
 - Assistive self-learning
 - Lost skills (long division?)
 - Is knowledge the ability for reflection or for problem solving?
- Science
 - Lab automation / hypothesis crunching
 - Is scientific advance about the results or about the knowledge creation?
- Politics
 - Simulating the 'average person' / data tell us all we need to know
 - The voice of the people or service to the people ?

**multidisciplinary
innovation
needed!**



TRUSTWORTHY GOVERNANCE

- Development of AI
 - The tech requirements
- Use of AI
 - The license to operate
- Context in which AI is developed and used
 - The rules of the game
- Global efforts
 - Sustainability
 - Inclusion and participation
 - Diversity
 - Distribution of benefits and costs
 - Agenda 2030 - SDGs

**political will and
innovation
needed!**



PROVABLE TECHNOLOGY

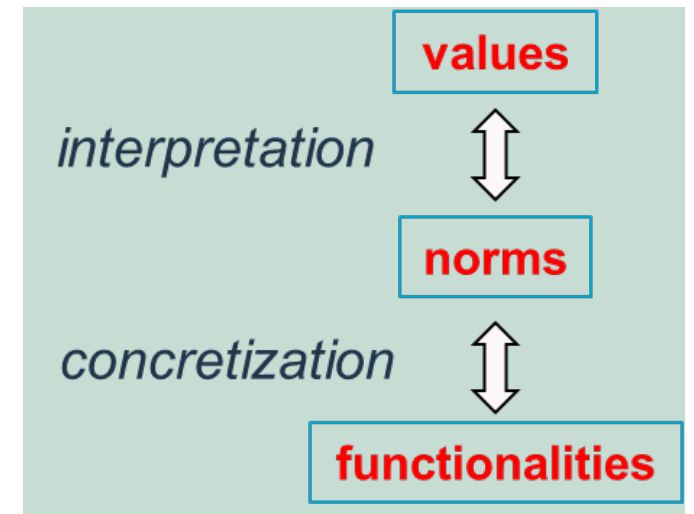
- Predictable
- Transparent
- Formally Verifiable
- Robust to Adversarial Attacks
- Generalizable
- Resilient
- Interpretable
- Data integrity
- Safe
- ...



**multidisciplinary
innovation
needed!**

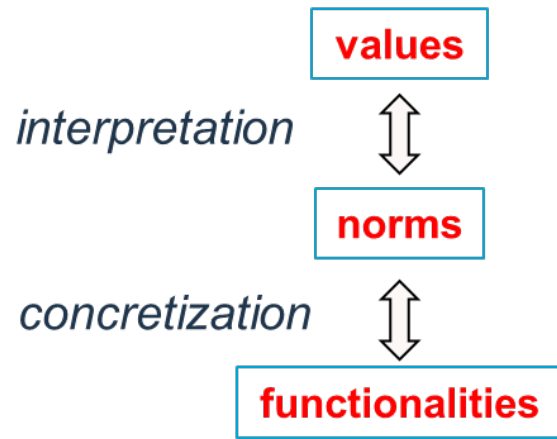
RESPONSIBLE AI IS ABOUT BEING EXPLICIT

- Design for Values
 - Legal and ethical aspects are not an add-on
- Regulation
 - External monitoring and control
 - Agreements, contracts, norms
- Design
 - Question your options and choices
 - Motivate your choices
 - Document your choices and options



DECISIONS MATTER!

Context matters!



safety

vehicles speed

users behavior



short term

long term

fairness

resources

opportunities



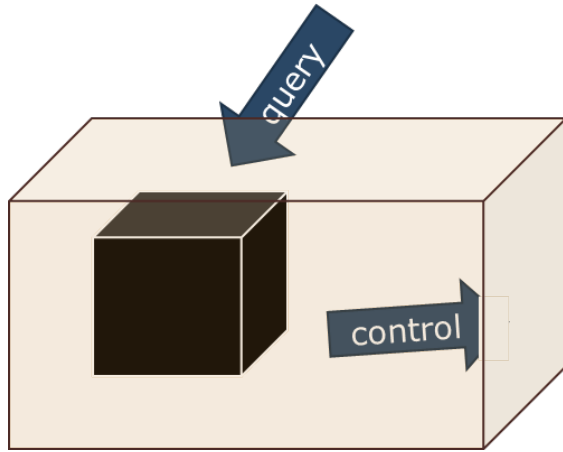
equality

equity



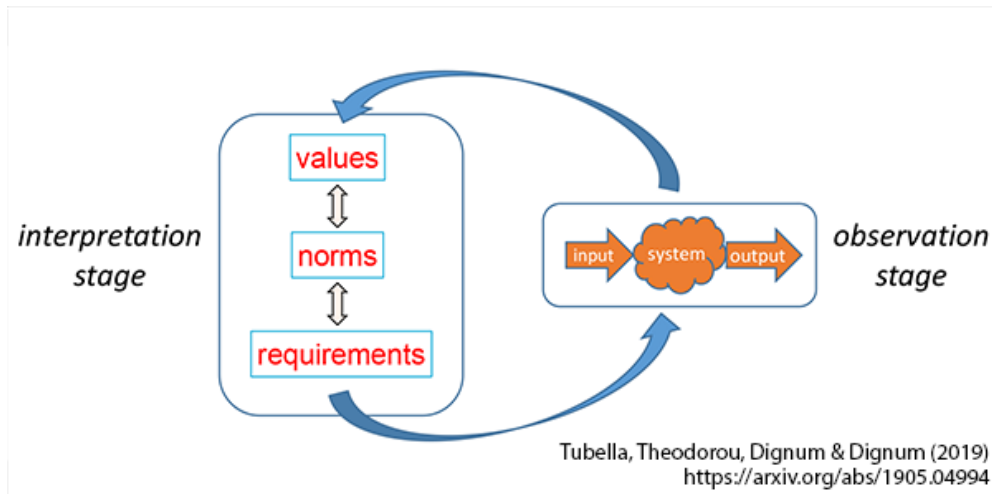
UMEÅ UNIVERSITY

GOVERNANCE - GLASS BOX



Desiderata for formal verification of ethical principles

- **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.



Tubella, A. A., Theodorou, A., Dignum, V., & Dignum, F. (2019). Governance by glass-box: Implementing transparent moral bounds for AI behaviour. *arXiv preprint arXiv:1905.04994*.



UMEÅ UNIVERSITY

DESIGN CRITERIA



- Internal independence
 - Cannot rely on internal structures
 - Must assume computational and organizational implementations
 - Validation



- Interaction independence
 - Must enable for different approaches
 - Human in/out/on the
 - Verifiability

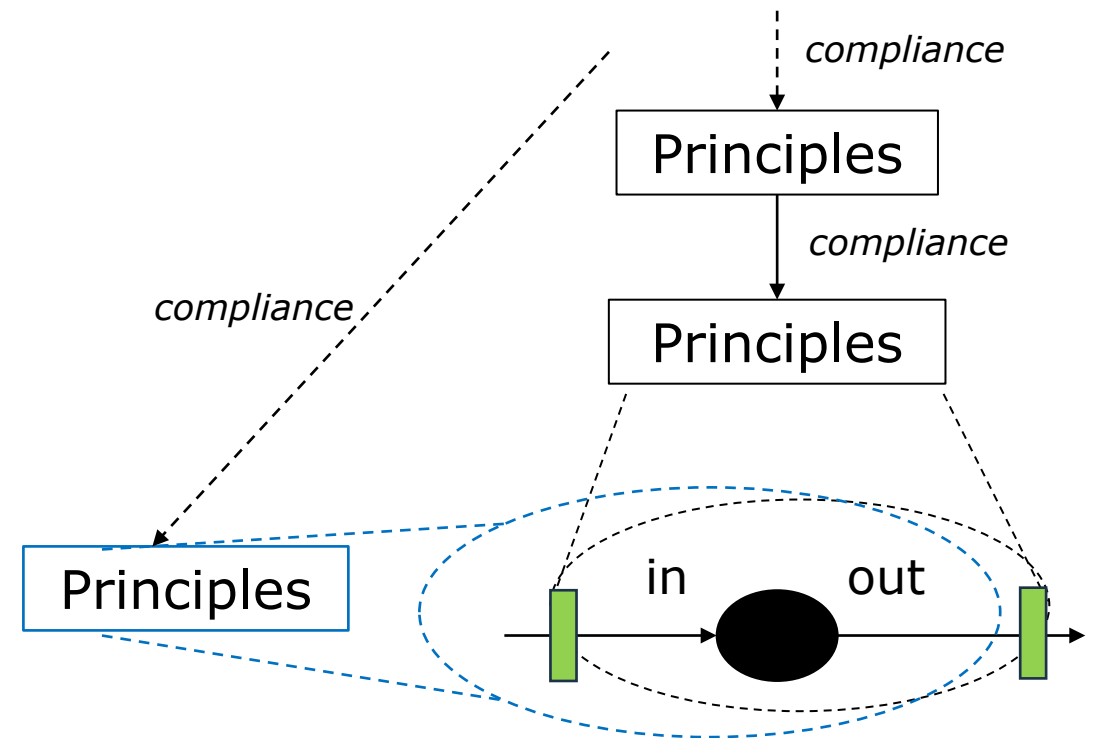


- Auditing independence
 - Computational / formal verification
 - Institutional arrangements
 - Contractual commitments



GOVERN AND VERIFY - GLASS BOXES

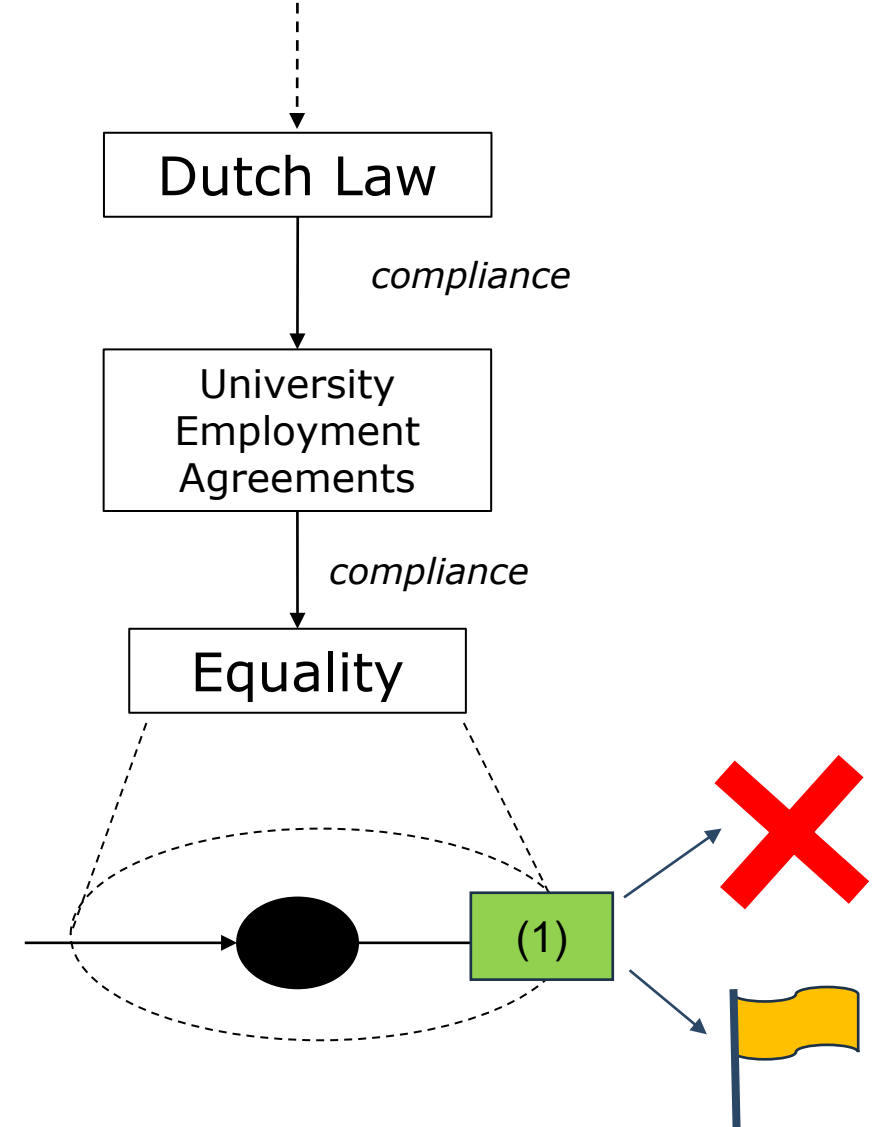
- Verify limits to action and decision
- Define the ethical borders
 - Formal
 - Monitoring input – output
- Governance
 - Monitor
 - “block” undesirable



EXAMPLE - FAIRNESS

- Value: Fairness
- Norm: Equality
- Implementation:
 - Output evaluation

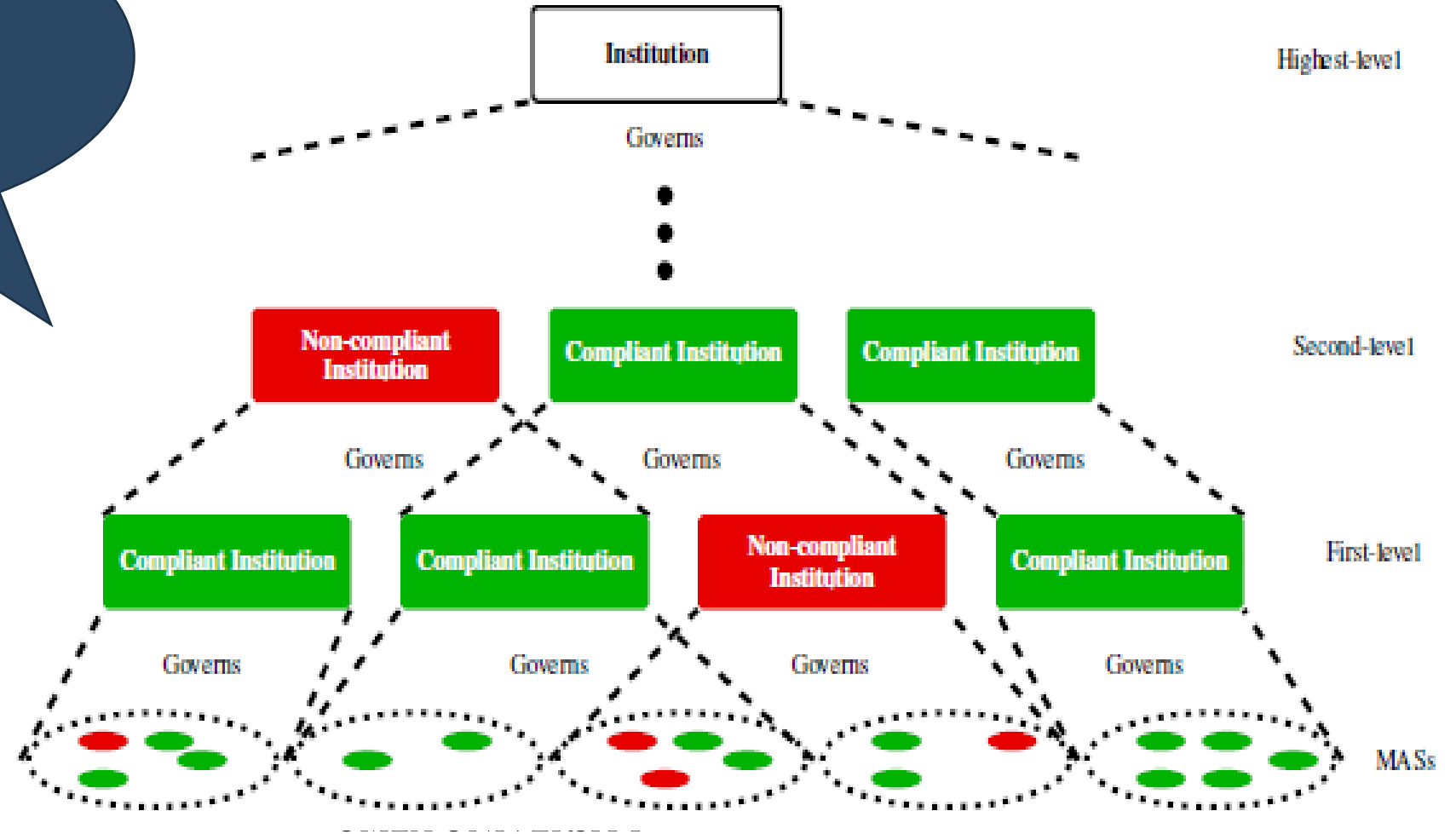
(1) $P(\text{job} \mid \text{female}) = P(\text{job} \mid \text{male})$
- Governance
 - Cut-off
 - Flag-out



GOVERNANCE TRANSPARENCY

We can also check consistency of supervising entities!

Increasingly Abstract Regulations



FORMALISING THE GLASS BOX: COUNTS AS

We will use two operators:

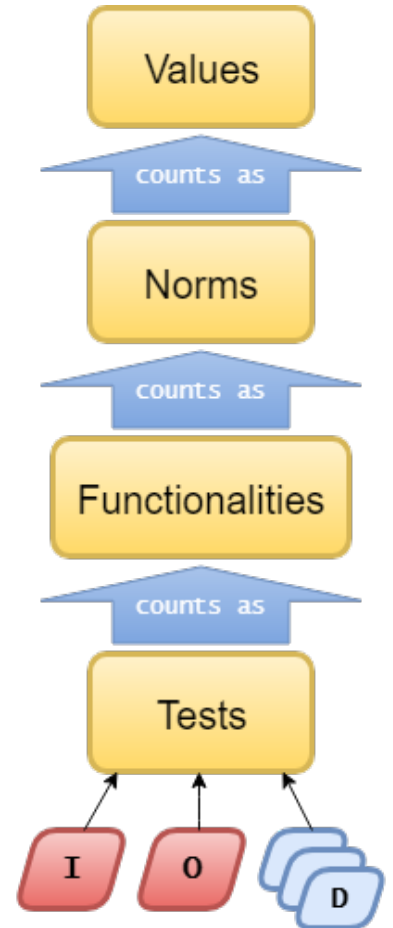
- “A counts-as B in context c” **by definition:** *constitutive* counts-as.
- “A counts-as B in context c”: *classificatory* counts-as.

These operators admit a formalisation in multi-modal logic.

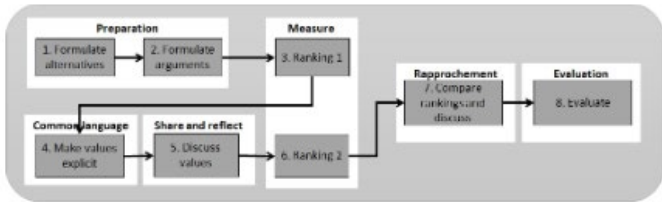
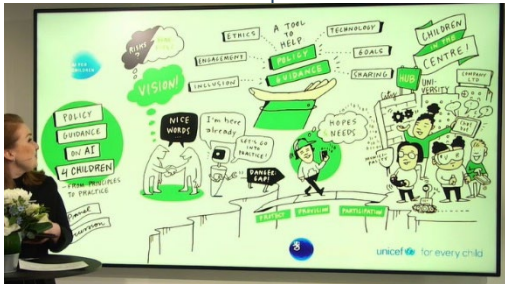
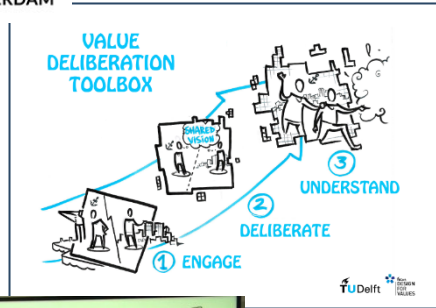
$$\gamma_1 \Rightarrow_c^{cl} \gamma_2 := [c](\gamma_1 \rightarrow \gamma_2)$$

$$\gamma_1 \Rightarrow_{c,\Gamma}^{co} \gamma_2 := [c]\Gamma \wedge [-c]\neg\Gamma \wedge \neg[u](\gamma_1 \rightarrow \gamma_2)$$

- Answering the question “**does A count as B in context c in the Glass Box?**” is equivalent to checking whether the implication $A \rightarrow B$ holds propositionally with the assumptions of c.
- This means that we can reason about what holds in the Glass Box in **reasonable time** (well within the reach of SAT-solvers and answer set programming approaches).



DESIGN FOR RESPONSIBILITY

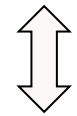


Verdiesen, I., Aler Tubella, A., & Dignum, V. (2021). Integrating Comprehensive Human Oversight in Drone Deployment: A Conceptual Framework Applied to the Case of Military Surveillance Drones. *Information*, 12(9), 385.

Pigmans, K., Aldewereld, H., Dignum, V., & Doorn, N. (2019). The role of value deliberation to improve stakeholder participation in issues of water governance. *Water Resources Management*, 33(12), 4067-4085.

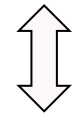
interpretation

values



norms

concretization



functionalities



UMEÅ UNIVERSITY

VIRGINIA DIGNUM; EMAIL: VIRGINIA@CS.UMU.SE - TWITTER: @VDIGNUM

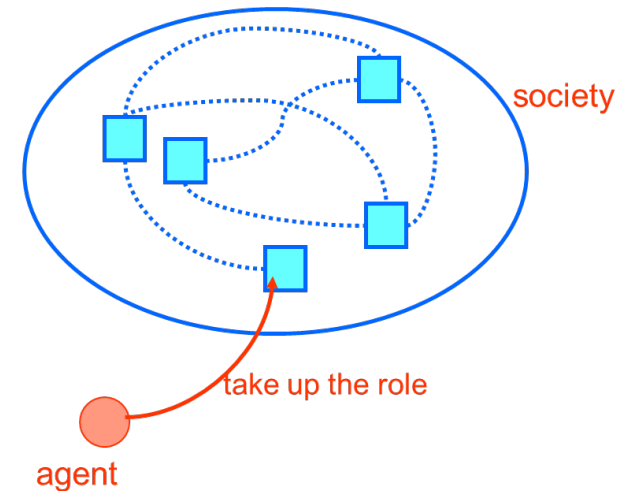
PROVABLE INTERACTION - THE OPERA APPROACH

Back to the future?! (PhD 2004)

- Agreement technologies
- Normative multi-agent organisations
- Logic of organisations

Distinguish society/collective aspects from individual issues

- Role: explicit, agreed, description of the part in the interaction
- Actor: individual capabilities, abilities, enhancement of role
 - AI/Robot
 - Human
 - Organisation
- Play: unique instance of the interaction of role-enacting actors

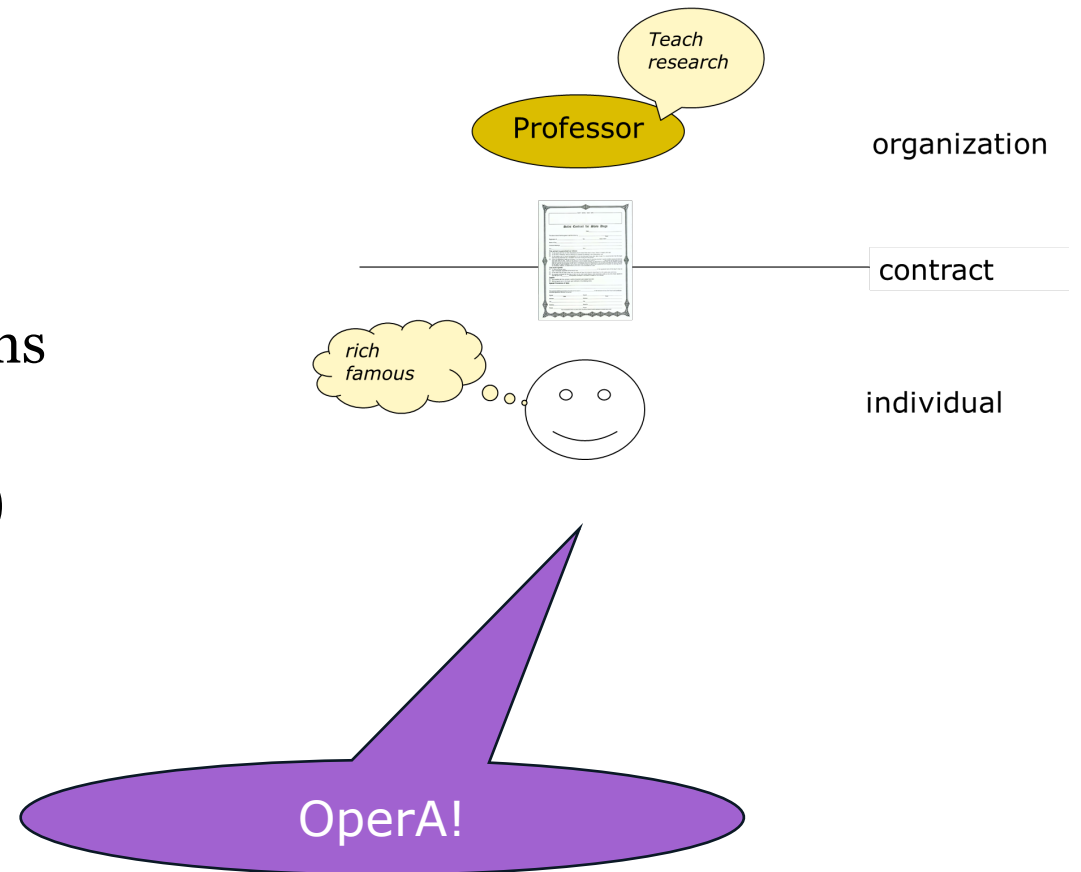


UMEÅ UNIVERSITY



TRANSPARENCY AS CONTRACT

- Contracts
 - Abstract from internal 'operation'
 - Support autonomy
 - Impose constraints and expectations
 - regulate behavior among agents
 - Require control / trust (3rd parties)



WHO IS RESPONSIBLE?

- “Program Committee members must return their reviews before the deadline”
 - Each PC member make a review
- “Project team member must file in a meeting report within 48 hours after the meeting”
 - One report for the whole team
- “Children below 16 are obliged to attend school”
 - Children must act, parents are blamed

	Individual Responsibility	Representative Responsibility	Collective Responsibility
Individual Actorship	$\forall x \in G : \{x\} \mathbf{D}_G^{\{x\}} \varphi$	$A \preccurlyeq R,$ $\forall x \in G : \{x\} \mathbf{D}_G^R \varphi$	$R = G,$ $\forall x \in G : \{x\} \mathbf{D}_G^R \varphi$
Representative Actorship	$A \preccurlyeq G,$ $\forall x \in G : {}^A \mathbf{D}_G^{\{x\}} \varphi$	$A \preccurlyeq G, A \preccurlyeq R :$ ${}^A \mathbf{D}_G^R \varphi$	$A \preccurlyeq G, R = G :$ ${}^A \mathbf{D}_G^R \varphi$
Collective Actorship	$A = G,$ $\forall x \in G : {}^A \mathbf{D}_G^{\{x\}} \varphi$	$A = G, A \preccurlyeq R :$ ${}^A \mathbf{D}_G^R \varphi$	$A = R = G :$ ${}^A \mathbf{D}_G^R \varphi$

CONTESTATION AI DECISIONS

- Formal language for integrating socio-ethical values into AI, detailing norms and requirements to better align AI with societal expectations.
- Feedback architecture to monitor AI-driven predictions and decisions against these norms and ethical standards.
- Framework for evaluating the contestability of black-box predictions, enhancing oversight and assessment capabilities for decisions by autonomous systems.



CONTESTATION FORMALISM

- Based on argumentation theory
- Types of contestation arguments:
 - T1: **values**. Eg. the decision is not fair.
 - T2: **norms**. Eg. people in my income bracket have been approved.
 - T3: **factual errors**. Eg. you have claimed that I am unemployed, but I actually am, here is employment record
 - T4: **instantiating errors**. Eg. you operationalized the norm wrong, you implemented the norm wrong, you selected the wrong norm.
 - T5: **misplaced counts-as**. Eg. you have assumed that my children are dependent, but they are grown up.
 - T6: **similarity**. Eg. you have said yes to a person who has the same properties as me.
 - T7: **counter examples**. Eg. you said that this was the only way to accomplish X, but here is another way to do it.



Arg1

Decision by the black-box system: *Bad customer*

Justification: LoanDuration > 13.50 -> PurposeOfLoan <= 7.50 -> Single <= 0.50 -> PurposeOfLoan <= 4.50 -> LoanDuration <= 42.00 -> Unemployed <= 0.50 -> LoanDuration > 22.50 -> Credit > 2393.50 -> Age > 46.50

AF_0 = ({arg}_0, {}_0)

Arg2

Client: The classification is unfair and possibly based on age discrimination.
Evidence: Different outcome for younger client in similar situation.
Request: Explanation of decision criteria and justification of age as a factor.

Arg3

Client: The classification has taken my Civil Status into account.
Evidence: Different outcome for single client in similar situation.
Request: Explanation of decision criteria and justification of civil status as a necessary attribute.

Arg4

Client: The loan requested for opening a business is more important than the one requested for buying a new car
Evidence: Different outcome for a client in similar situation when buying a new car.
Request: Explanation of decision criteria and justification of the importance of the purpose of the loan.

AF_1 = ({Arg1, Arg2, Arg3, Arg4}_1, {(Arg2, Arg1), (Arg3, Arg1), (Arg4, Arg1)}_1)

Arg5

BB: Justification: Disclosure of internal policies that justify the decision criteria, including any age-related considerations.

Arg6

BB: Justification: Information on how the decision path is determined, including the role of the civil status in the algorithm.

Arg7

BB: Justification: Information on how the decision path is determined, including the role of the purpose of the loan in the algorithm.

AF_2 = ({Arg1, Arg2, Arg3, Arg4, Arg5, Arg6, Arg7}_1, {(Arg2, Arg1), (Arg3, Arg1), (Arg4, Arg1), (Arg5, Arg2), (Arg6, Arg3), (Arg7, Arg4)}_2)

Arg8

Client: Civil status should not solely determine loan approval, especially given:
Factors Favoring Client: Homeownership, Employment.
Client's Claim: The decision represents discrimination based on civil status. This attribute not should be used for the decision making.

Arg9

Client: The purpose of the loan should not solely determine loan approval, especially given:
Factors Favoring Client: Homeownership, Employment, Specific use of loan for a business venture.
Client's Claim: The decision represents discrimination based on the purpose of the loan.

AF_3 = ({Arg1, Arg2, Arg3, Arg4, Arg5, Arg6, Arg7, Arg8, Arg9}_1, {(Arg2, Arg1), (Arg3, Arg1), (Arg4, Arg1), (Arg5, Arg2), (Arg6, Arg3), (Arg7, Arg4), (Arg8, Arg6), (Arg9, Arg7)}_3)

Arg10

BB: Justification: Disclosure of internal policies that justify the decision criteria, designed to protect the financial stability.

AF_4 = ({Arg1, Arg2, Arg3, Arg4, Arg5, Arg6, Arg7, Arg8, Arg9, Arg10}_1, {(Arg2, Arg1), (Arg3, Arg1), (Arg4, Arg1), (Arg5, Arg2), (Arg6, Arg3), (Arg7, Arg4), (Arg8, Arg6), (Arg9, Arg7), (Arg9, Arg10)}_4)

FORMAL APPROACHES IN RESPONSIBLE AI

- **Ensuring Trust and Accountability**
 - Formal approaches enable predictable, transparent, and verifiable AI systems.
- **Advancing Governance and Innovation**
 - Formal methods facilitate the integration of comprehensive human oversight and the evaluation of AI's ethical impacts, and support contestability, explanation and comparisons
- **Promoting Global Standards and Equity**
 - Formal approaches support the creation of universally applicable standards that respect diversity and ensure equity.
- **Addressing Real-world Complexity**
 - Formal methods allow for rigorous testing and adaptation of AI systems to various real-world conditions, enhancing their resilience and effectiveness.



THANK YOU!

virginia@cs.umu.se

Twitter: @vdignum



UMEÅ UNIVERSITY