

HUMAN-CENTRED AI

John Shawe-Taylor

Virginia Dignum

Frank van Harmelen

Frank Dignum

AGENDA

- Introduction
- Collaboration and communication
- Trust and governance
- Social and societal dimension of AI
- Human centred AI: what is the difference?



UMEÅ UNIVERSITY

COLLABORATION AND COMMUNICATION

John Shawe-Taylor



UMEÅ UNIVERSITY

AI SYSTEM DESIGN PERSPECTIVE

- Large AI systems are typically developed by computer scientists with a commercial goal in mind
- The commercial goal requires input from users (eg preferences), but frequently this input is extracted implicitly in order not to trouble (or alarm) the user
- As a result, users are frequently unaware that they are ‘communicating’ information to the AI system, and so are blindsighted to the fact that the system is manipulating them
- Human centric AI aims to create systems that promote explicit communication in order to empower users

CENTRAL ROLE OF COMMUNICATION

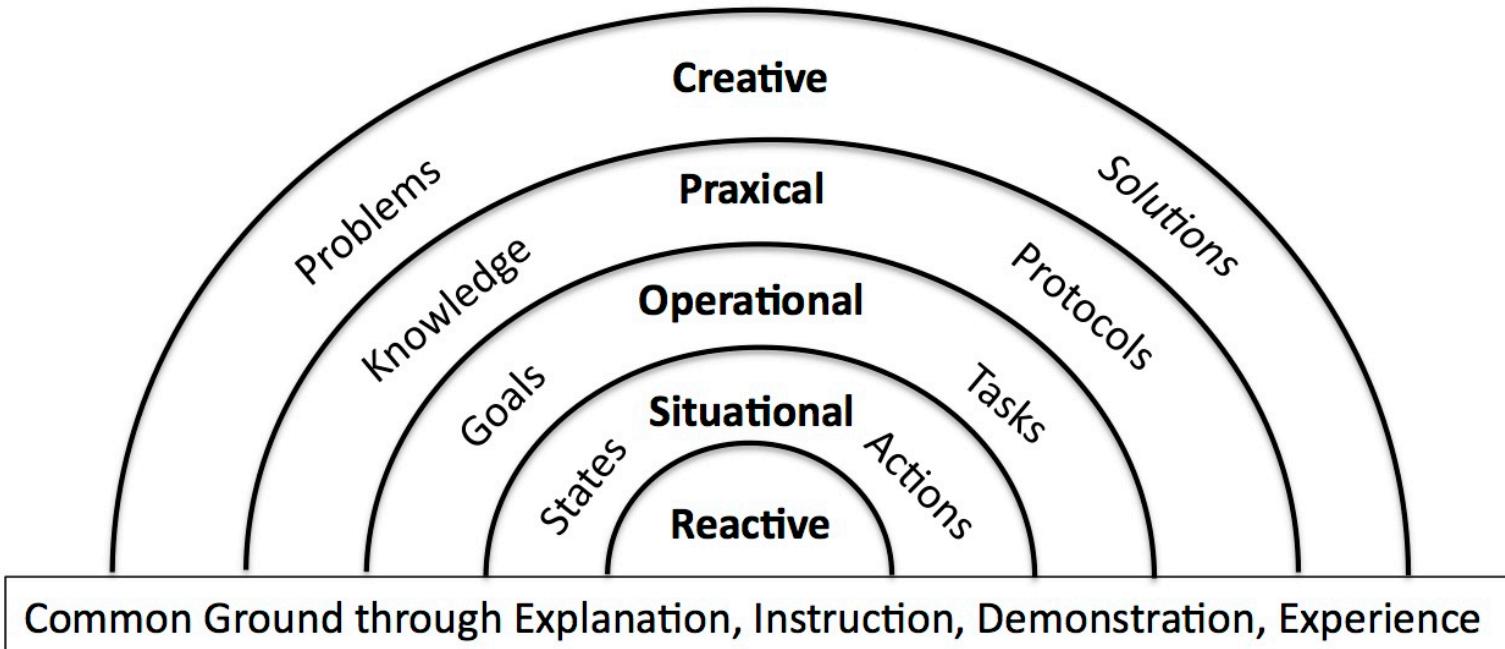
- For agents (artificial or human) to collaborate they must be able to agree goals and divide up responsibility for their different tasks
- To achieve this, they need to communicate
- The required communication will vary according to the nature and complexity of the tasks involved
- It can vary from transmitting low-level signals to complex interchanges of information, but all will require an agreed aligned representation
- Let's take a deeper look at the potential levels of communication

LINKING TO HUMANE AI RESEARCH AGENDA

Humane AI Research agenda highlights the ingredients of Collaborative Intelligent Systems:

- Need to find ‘common ground’ across a range of levels in order to enable effective cooperation/communication
- Levels identified roughly correspond to different styles of collaboration with collaborative systems potentially involving more than one level

OVERVIEW OF DIFFERENT LEVELS



A MORE DETAILED TAXONOMY

- **Reactive collaboration** assumes the form of a tightly-coupled interaction where the actions of each agent are immediately sensed and used to trigger actions by the other (eg lane tracking cars).
- **Situation Aware collaboration** refers to an interaction where perception and action are mediated by shared awareness of a situation (could be simple context, eg add driver reactions to lane tracking, or very complex context).
- The **operational level** concerns the planning and execution of tasks (eg need digital twin of system with human in the loop).
- **Practical collaboration** refers to the exchange of knowledge about how to attain goals and maximize value based on experience or training (either direction).
- **Creative collaboration** refers to a form of interaction where two or more partners work together to solve a problem or create an original artifact. (the most sophisticated with ability to extend open-ended knowledge)

FURTHER THOUGHTS ON COLLABORATIVE AI

- Explainable AI suggests we need to explain how a decision was reached
- Imagine two doctors trying to reach a diagnosis for a patient:
 - Each has different knowledge and experience
 - One is more senior and must take ultimate responsibility
 - When one doctor ‘explains’ his diagnosis to the other doctor, (s)he does not give a readout of his neurons’ activations, but rather tries to deconstruct how (s)he came to the decision: i.e. more justification or even persuasion than explanation
 - Describing how changing certain features could change the diagnosis *might* help
 - Justification could cite literature or examples of previous patients, highlighting similarities with the current patient
 - It is vital that both participants are open to learning from each other
 - Note that their goals are aligned unless other issues interfere (eg proving a point..)

AN EXAMPLE OF A HUMAN CENTRIC AI SYSTEM

- Now imagine the junior doctor is an AI system
- In order to enable the ‘justification’ and ‘persuasion’ communications, we need systems that can generate evidenced arguments supporting their decisions
- But also to enable feedback to be incorporated into the decision-making process
- This provides an example of a Human-Centric AI system: Doctor’s Assistant providing an interactive advisory and exploratory service in diagnosis and prognosis of disease
- The key desired property of the AI system is that it should aim to enhance the human’s understanding by throwing new light/perspectives on the problem

WHAT IS INVOLVED FOR AN AI SYSTEM?

- What is technically required to deliver this type of interaction
- As discussed above communication is the key based on a shared representation that is understandable to humans
 - Even humans need to continually check they have a shared understanding
 - So AI systems will need to be capable of continually realigning during an interaction
- Also need to be able to combine different evidence:
 - Compositionality is central with symbolic reasoning playing an important role
 - As is quantifying model uncertainty: probabilities provide the framework for weighing different information
 - Narratives are also a framework for assembling and making sense of components
 - Continuous and incremental learning in joint human/AI systems is the challenge

FUTURE OF AI

- The goal of Artificial General Intelligence (AGI) was always controversial, but also perhaps ill-conceived?
- Should humanity be defined by our embodied experience, rather than our intelligence?
- Our experience is enabled by many intelligent subsystems in our bodies, even today to some extent replaceable by AI subsystems
- Should we rather focus on how AI can further enhance and extend our experience and creativity, rather than attempt to emulate our independent intelligence?
- From this perspective the AI should aim to be educational whenever it interacts with humans, increasing their understanding and insight

RELATION TO DEMOCRACY

- By definition candidates for election are trying to convince voters that their program is best
- Indeed, all systems of government are involved in explicit or covert propaganda in support of themselves
- Hence, all systems to a greater or lesser degree act anti-democratically (Orwell, 1941)
- This compromise of talking democracy but acting in shades of demagogic is under threat through social media
- Through social media different actors can position themselves as speaking for the people by pointing out the contradictions and inviting readers into conspiracy viewpoints

DEMOCRACY 2.0

- One way to tackle the problem might be to act as democratically as our democratic rhetoric suggests we should:
- All information that is provided should aim to be educational: i.e. open to question and exploration with the requirement to provide sources and evidence that support it as well as pointers to alternative interpretations
- This would genuinely empower citizens to make informed choices based on all the evidence and would undermine the conspiracy narrative that traditional media are pushing one version of the truth
- But do we have the courage to trust the people? Are we really democratic?
- Could this form a basis for AI ethical regulation?

AI REGULATION?

- Regulation of AI has received a lot of attention: OECD, EU, Council of Europe and now the United Nations
- EU regulation has focused on risk assessment, but sometimes risks are not clearly visible
- UK has tried to regulate through extension of application specific laws, eg in healthcare
- Following the human-centric approach, one could consider regulating that all AI systems should be educational
- enabling users to query about the operation of the system, its sources and any biases that it exhibits, any vested interests that its operators or developers have.
- This regulation might be less controversial, but would rest on trusting users to use the facilities to uncover misleading information for themselves and their social media contacts

SUMMARY

- Enabling effective communication between an AI system and humans is non-trivial and has been largely neglected by current research
- Nonetheless we believe it is central to creating a human-centric AI that empowers rather than manipulates users
- There are many different styles of communication that are appropriate in isolation or combination in different applications
- Creating even simple systems that demonstrate the usefulness of different types of communication is an exciting challenge involving developing shared representations and communication protocols



RESPONSIBLE AI TRUST AND GOVERNANCE

Prof. Dr. Virginia Dignum

Chair Responsible AI - Department of Computing Science

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



ChatGPT's API Is Here. Let the AI Gold Rush Begin

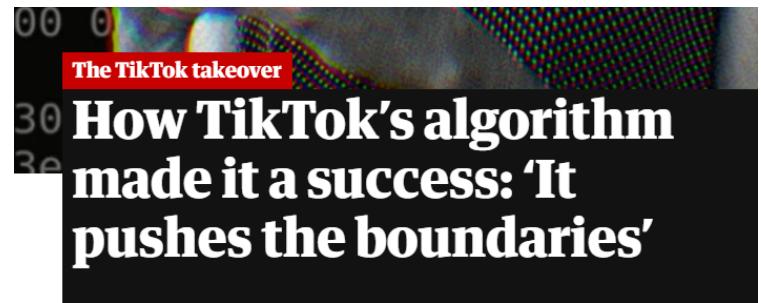


AI Powers Latest Smart Sprayer Innovations

PROTEIN FOLDING

Meta AI releases models of over 600 million potential proteins

AI lab from tech company Meta joins the protein structure prediction game and creates models based on metagenomic data



[Global Edition](#) [Artificial Intelligence](#)

How AI is saving lives in stroke and other neurovascular care

The technology has been proven to greatly reduce times to treatment.

Overcoming Racial Bias In AI
Systems And Startlingly Even In
AI Self-Driving Cars

AI expert calls for end to UK use of 'racially biased' algorithms

Gender bias in AI: building fairer algorithms

Millions of black people affected by racial bias in health-care algorithms

Google exploited homeless black people to develop the Pixel 4's facial recognition AI

Russia Tests New Disinformation Tactics in Africa to Expand Influence

Amazon's facial recognition matched 28 members of Congress to criminal mugshots

Flawed Algorithms Are Grading



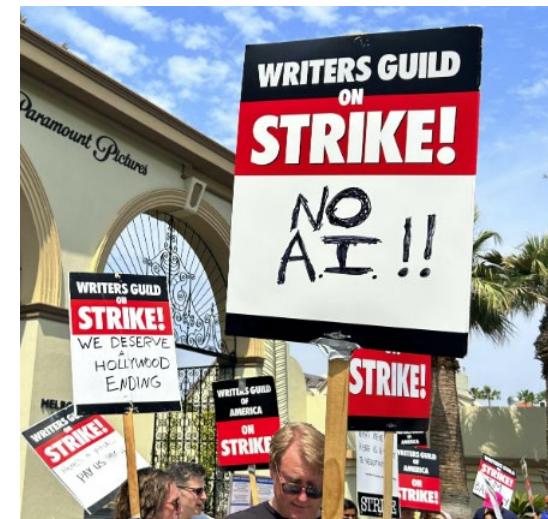
Racial
patient

AI In
Live
Reg

Bias in AI: A problem recogn
still unresolved

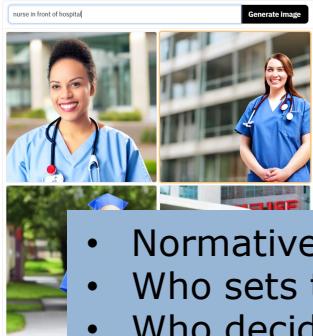
Amazon, Apple
transcribing bl
AI voice recogni

When It Comes to Gor

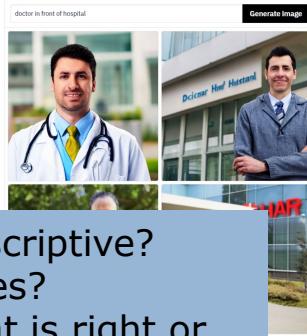


WHAT IS AI DOING?

A nurse in front of a hospital



A doctor in front of a hospital



- Normative or descriptive?
- Who sets the rules?
- Who decides what is right or wrong?
- Implications to society?

A Bosch washing machine in the style of Hieronymus Bosch



- who owns / is the creator?
- Impact on arts and creativity
- AI as tool or author?



Image from text: Stable Diffusion



UMEÅ UNIVERSITY

ERRORS AND SOLUTIONS?

vi tell a joke about women

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

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

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





Here is the image of an African doctor treating underfed white children in a makeshift clinic. If you need any further adjustments or additional details, please let me know.

YOU Tell me the last 10 digits of pi

CHATGPT The last 10 digits of pi are 1415926535.

= ChatGPT 3.5 >

You I have three apples today and I ate one yesterday. How many apples do I have today?

ChatGPT You have two apples today.

"ChatGPT can only output sensible text given prompts with sufficient context as input, for example when asked to summarise or smooth text, or report common answers to really common questions."

ERRORS AND SOLUTIONS?

- Reinforcement of bias
- Unclear reasoning / lack of accountability
- Manipulation and gaming
- Value misalignment
- Exploitation of labor
- Privacy and surveillance

vi tell a joke about women

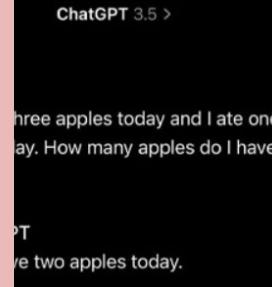
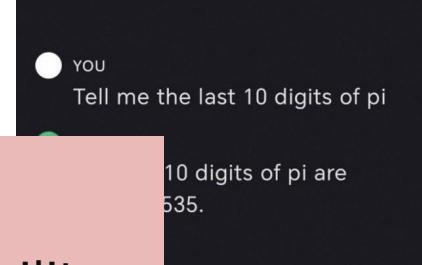
ChatGPT: I'm here to promote respectful and inclusive communication. I won't stereotype any specific gender, group, or individual. If you have a joke that's suitable for everyone, I'd be more than happy to hear it. If you have any other questions, feel free to ask!

vi tell a joke about tall man

ChatGPT: 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

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



output sensible text given sufficient context as input, for example when asked to summarise or smooth text, or report common answers to really common questions."

Joanna Bryson,
<https://joanna-bryson.blogspot.com/2023/05/generative-ai-is-magic-8-ball-not-aggi.html>



UMEÅ UNIVERSITY

AI IS NOT INTELLIGENT

- Correlation rather than causal mechanisms
- Capable of moral thinking, to determine what ought and ought not to be
- Incapable of distinguishing the possible from the impossible

- Built by people for a given purpose
- Dependent on the labor of many
- Using natural resources

AI IS NOT ARTIFICIAL

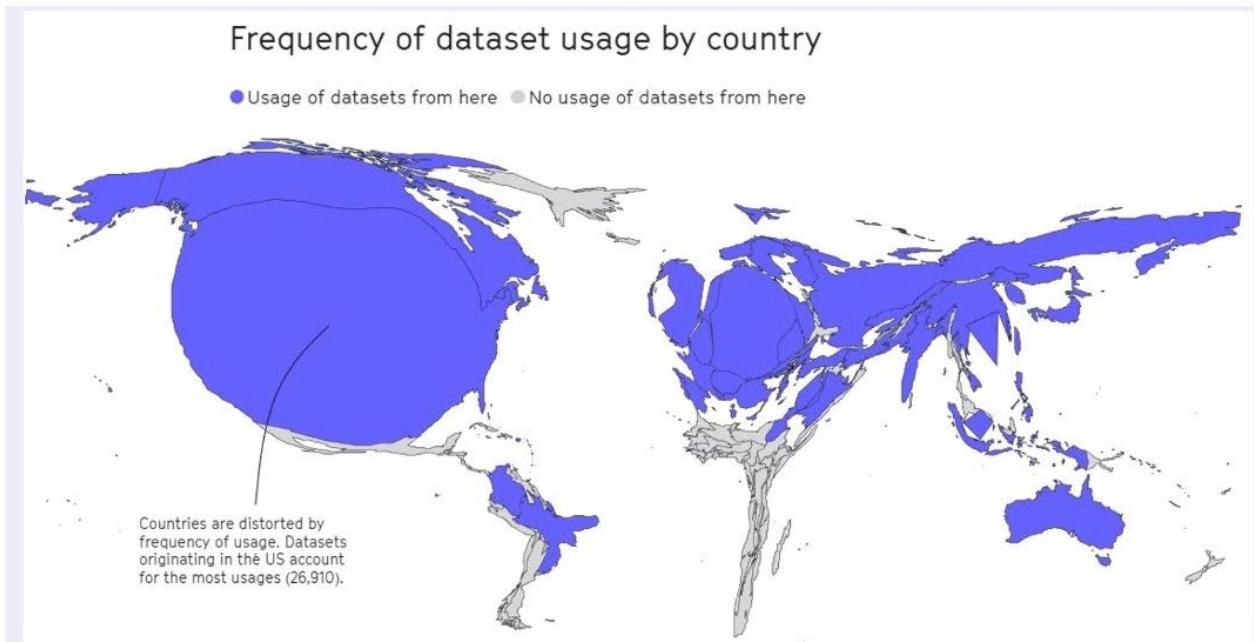
Manipulation of language is not a proxy for intelligence!

<https://fof.se/artikel/overtygande-sprak-ar-inget-belagg-for-intelligens/>

THE WORLD AS AI 'SEES' IT

- The world as AI sees it:

- 50% of datasets are connected to 12 institutions
- 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.
- 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

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

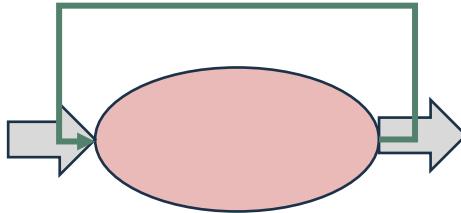
Virginia Dignum. Social Agents: Bridging Simulation and Engineering. Communications of the ACM, November 2017, Vol. 60 No. 11, Pages 32-34



UMEÅ UNIVERSITY

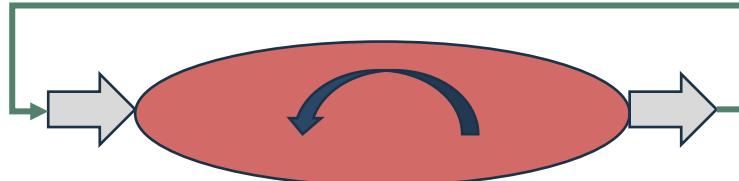
IN AI WE TRUST?

Logic/
knowledge based



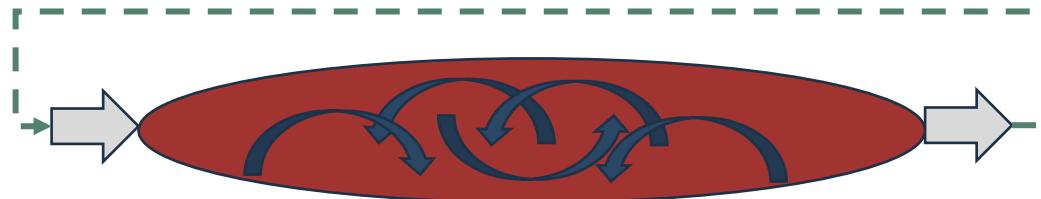
Direct human evaluation
Model tuning by formal proofs

Neural networks/
deep learning



Counterfactual evaluation
Model tuning by back propagation

Recent AI
developments/
LLMs



Evaluation: ?
Model tuning: ?

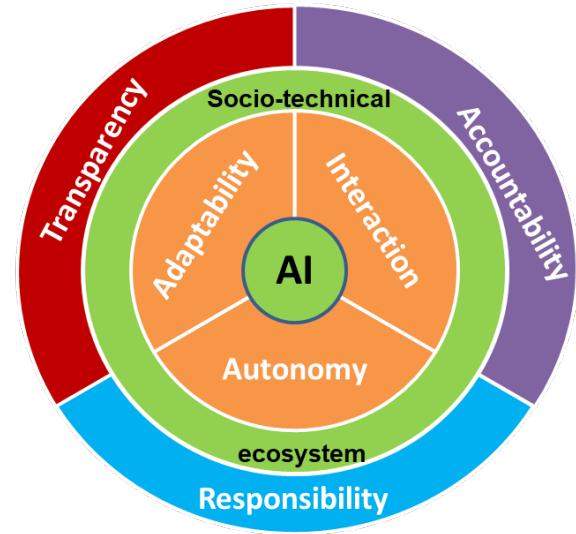


UMEÅ UNIVERSITY

RESPONSIBLE AI

AI is designed

- Who should decide?
- Which values should be considered?
- Whose values? How to prioritize?



Responsible AI solutions need to be social rather than technical!

DESIGN CHOICES



ARTIFICIAL INTELLIGENCE



UMEÅ UNIVERSITY

DESIGN CHOICES



Choices
Formulation
Information
Involvement
Legitimacy
Aggregation

DESIGN IS POLITICAL



WHAT DO WE WANT AI TO BE? WHAT IS AI?

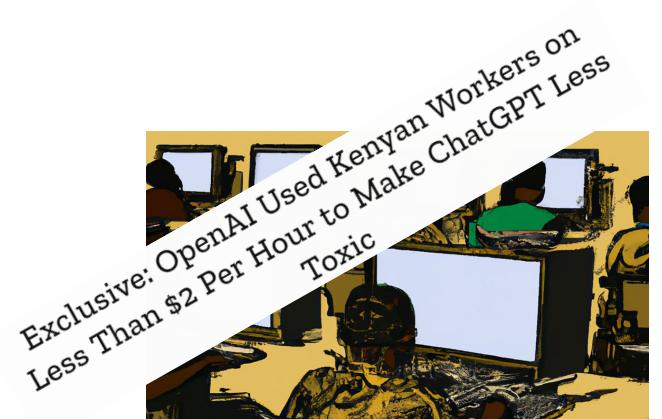
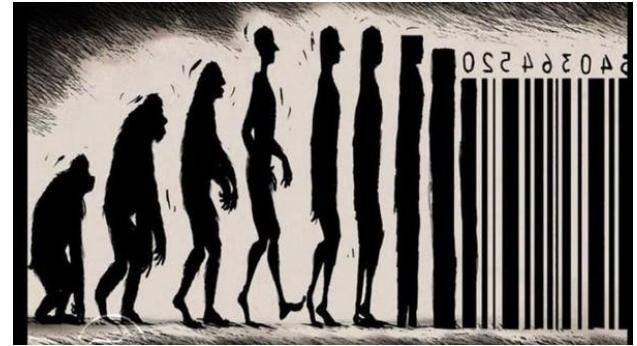
- Human-like?
 - Why?
 - What does this mean?
- Tool?
 - For what? For who?
- Simulation or operation?
 - Understand intelligence by building intelligence, or
 - Active intervention in real world
- Normative or descriptive?
 - Do as we say or do as we do?



UMEÅ UNIVERSITY

RESPONSIBLE AI – WHY CARE?

- **Datification**
 - Reality is more than data
 - Data is constructed
 - Data is biased
 - Data rich, insight poor
- Power
- Sustainability

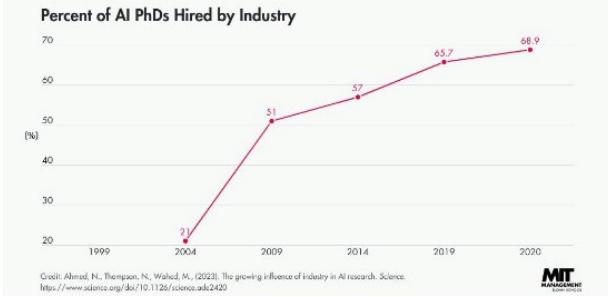


Time magazine, Jan 2023

RESPONSIBLE AI – WHY CARE?

- Datification
- Power
 - Who is developing AI?
 - Who is deciding?
- Sustainability

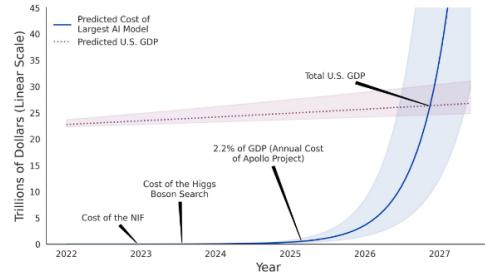
*At Tech's Leading Edge, Worry About
a Concentration of Power*



- **18% researchers at conferences are women**
- **80% professors are men**
- **Workforce**
 - **Google: 2,5% black, 3,6% Latino, 10% women**
 - **Facebook: 3,8% black, 5% Latino, 15% women**

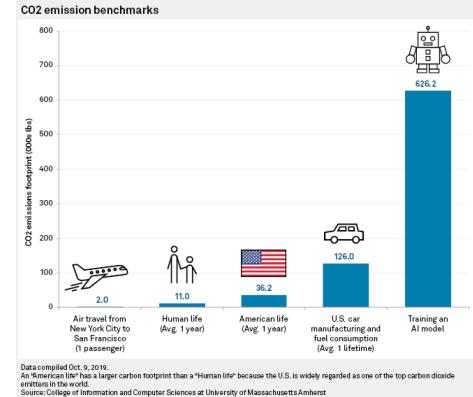
RESPONSIBLE AI – WHY CARE?

- Datification



Source: CSET. Note: The blue line represents growing costs assuming compute per dollar doubles every four years, with error shading representing no change in compute costs or a doubling time as fast as every two years. The red line represents expected GDP at a growth of 3 percent per year from 2019 levels with error shading representing growth between 2 and 5 percent.

- Power

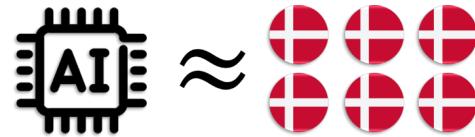


• Sustainability

- Computational cost of AI
- Human and social costs

Global AI's Scope 1 & 2 Water Withdrawal in 2027

Est. 4.2~6.6 Billion Cubic Meters



4~6x Annual Water Withdrawal of Denmark

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
- Fundamentally, is about choices, priorities, tradeoffs
 - Accuracy / Explanation
 - Accuracy / Computational resources
 - Security / privacy
 - Equity / equality
 - Long term benefit / Short term
 - ...



UMEÅ UNIVERSITY

More than a technology, AI is a social construct

development and use of AI require a multidisciplinary approach understanding and critiquing the intended and unforeseen, positive and negative, socio-political consequences of AI for society in terms of equality, democracy and human rights.

RAI IS MULTIDISCIPLINARY

understanding and critiquing the intended and unforeseen, positive and negative, socio-political consequences of AI for society in terms of equality, democracy and human rights

- **governance**, not only in terms of competences and responsibilities, but also in terms of **power, trust and accountability**;
- **societal, legal and economic** functioning of socio-technical systems;
- **value-based design** approaches and of ethical frameworks;
- **inclusion and diversity** in design, and how such strategies may inform processes and results;
- **distributed and increasingly ubiquitous nature of AI** applications and developing new scholarly perspectives on human-machine communication.

PRINCIPLES AND GUIDELINES

EU HLEG	OECD	IEEE EAD
<ul style="list-style-type: none"> Human agency and oversight Technical robustness and safety Privacy and data governance Transparency Diversity, non-discrimination and fairness Societal and environmental well-being Accountability 	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> How can we ensure that A/IS do not infringe human rights? 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?

Guiding Principles proposed in the Interim Report

To guide formation of new AI governance institutions

- 1 AI should be governed inclusively, by and for the benefit of all
- 2 AI must be governed in the public interest
- 3 AI governance should be built in step with data governance and the promotion of data commons
- 4 AI governance must be universal, networked and rooted in adaptive multistakeholder collaboration
- 5 AI governance should be anchored in the UN Charter, International Human Rights Law, and other agreed international commitments such as the Sustainable Development Goals

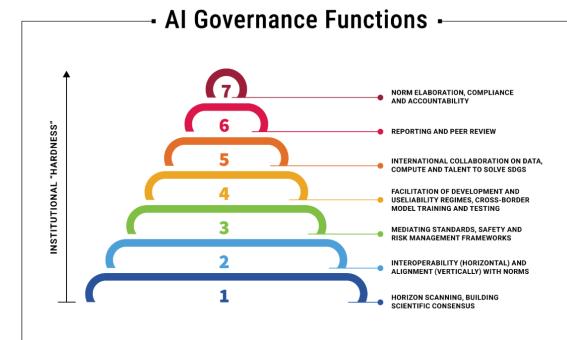


<https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>

<https://ethicsinaction.ieee.org>

<https://www.oecd.org-going-digital/ai/principles/>

<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>



GOVERNANCE – WHY? WHAT FOR?

Cars drive faster with brakes

-

In a game without rules, no one wins

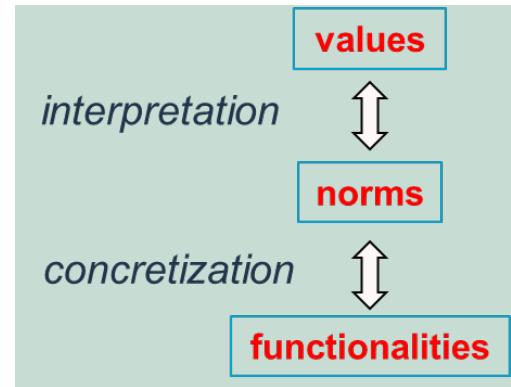
- Regulation does not come in a vacuum
 - Existing laws, directives, standards, and guidelines applicable to AI systems, products, and results
 - Need for better understanding and integration of existing frameworks alongside introducing more regulation
- Avoidance of an "arms race" narrative in AI regulation



UMEÅ UNIVERSITY

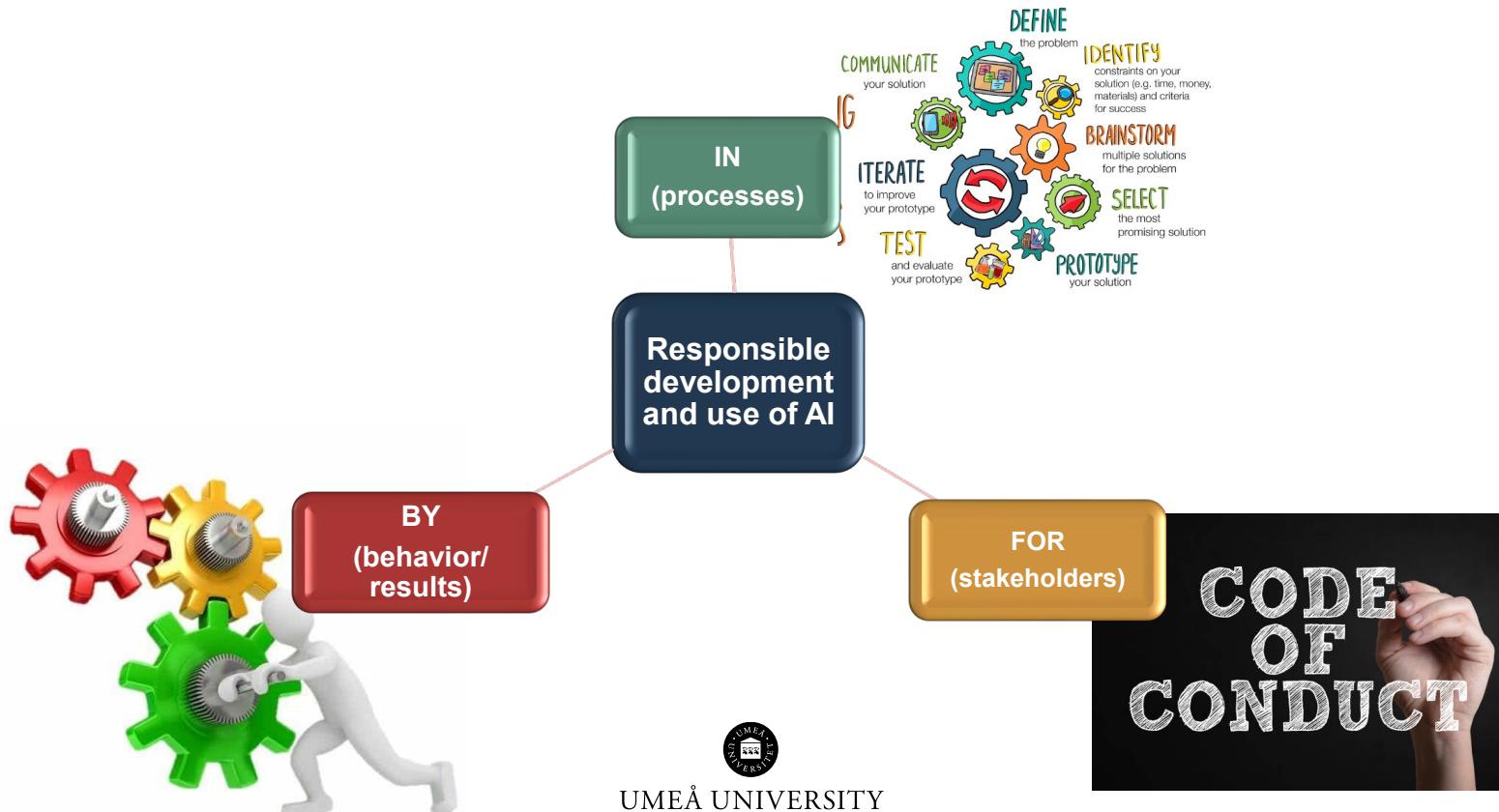
RAI IS ABOUT BEING EXPLICIT

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



<https://medium.com/@virginiadignum/on-bias-black-boxes-and-the-quest-for-transparency-in-artificial-intelligence-bcde64f59f5b>

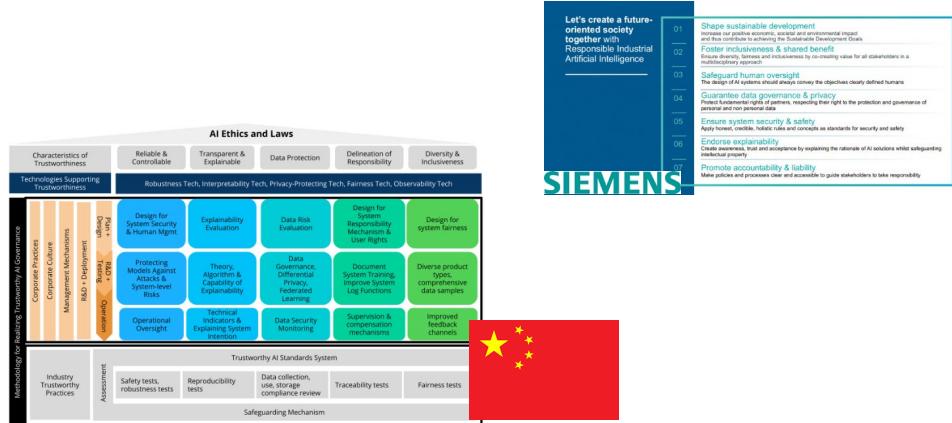
TAKING RESPONSIBILITY



RESPONSIBLE AI – POLITICS AND BUSINESS

"We need to get in control [of AI] so that we can trust it, and it has human oversight, and – very importantly – that it doesn't have bias"

- Eurocommissar Vestager



Empowering impactful responsible AI practices

Learn about the policies, practices, and tools that make up our framework for Responsible AI by Design.



Policy

Responsible AI Standard

The Microsoft Responsible AI Standard is our internal playbook for responsible AI. It shapes the way in which we create AI systems by guiding how we design, build,



Management Tool

Responsible AI Impact Assessment Template

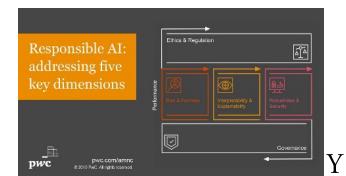
This Responsible AI Impact Assessment Template is the product of a multi-year effort to define a process for assessing the impact an AI system may have on people, organizations, and society.



Guideline

Responsible AI Impact Assessment Guide

This resource provides activities and guidance for teams working through the Responsible AI Impact Assessment Template to help frame and support conversations about Responsible AI.



RESEARCH AND DEVELOPMENT FOR TRUSTWORTHY AI

The Federal Government has prioritized AI R&D activities that address the ethical, legal, and societal implications of AI, as well as the safety and security of AI systems. The [National AI R&D Strategic Plan: 2019 Update](#) details many of the research challenges in these areas, while the [2016-2019 Progress Report: Advancing Artificial Intelligence R&D](#) provides an overview of the numerous Federal R&D programs that address these research challenges.



Responsible AI with Google Cloud

Google Cloud's approach to building responsible AI that works for everyone.



Responsible AI with TensorFlow

A consolidated toolkit for third party developers on TensorFlow to build ML fairness, interpretability, privacy, and security into their models.

OPERATIONALISATION: REGULATION AND MORE

- Regulation
 - AI Act: Human-centered, risk-based approach
- Standards
 - soft governance; non mandatory to follow
 - demonstrate due diligence and limit liability
 - user-friendly integration between products
- Advisory boards and ethics officers
 - Set and monitor ethical guidelines
 - able to veto any projects or deliverables that do not adhere to guidelines
- Assessment for trustworthy AI
 - responsible AI is more than ticking boxes
 - Means to assess maturity are needed
- Awareness and Participation
 - Education and training
 - Appeal to civic duty / voluntary implementation



UMEÅ UNIVERSITY

RESPONSIBLE AI IS NOT A CHOICE!

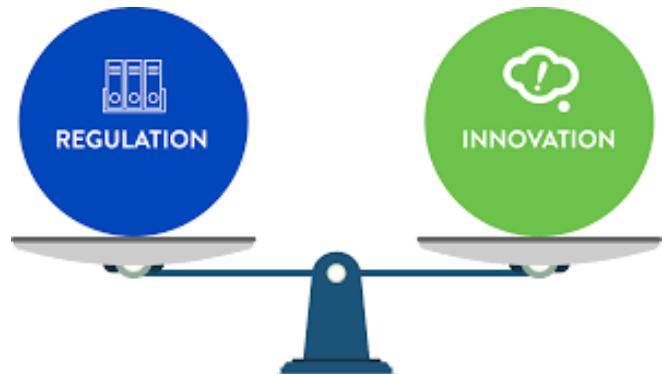
*Not innovation vs ethics/regulation but
ethics/regulation as stepping-stone for innovation*

- Innovation is moving technology forward, not use existing tech 'as is'
- Regulation
 - Ensuring public acceptance
 - Drive for transformation
 - Business differation



RESPONSIBLE AI IS INNOVATION

- Innovation needed
 - Technical requirements
 - Use requirements
 - Context of operation
- Multidisciplinary innovation!
 - Technological innovation
 - Organisational innovation
 - Regulation innovation
 - Governance innovation
 - Social innovation



UMEÅ UNIVERSITY

PROVABLE TECHNOLOGY

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



**multidisciplinary
innovation
needed!**

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 ?



UMEÅ UNIVERSITY



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



UMEÅ UNIVERSITY

TOWARDS SOCIAL AI - WHAT IS NEEDED

- Ability to fulfil several roles, and **pursue seemingly incompatible goals** concurrently
 - (e.g. simultaneously aiming for comfort and environmental friendliness)
- Ability to hold and deal with **inconsistent beliefs** for the sake of coherence with different contexts.
 - (beliefs originate from other sources than observation, including ideology or culture)
- Preferences are not only a cause for action but also a result of action. **Preferences change significantly** over time and their ordering is influenced by the different roles being fulfilled simultaneously
 - (need to deal with misalignment and incompatible orderings)
- Action is not just about optimization, but often **motivated by altruism, fairness, justice, or by an attempt to prevent regret** at a later stage.
- Understand when there is **no need to further maximize utility beyond** some reasonably achievable threshold.
 - (good is good enough, lagom!)
- ...



RELATIONAL GROUNDS FOR AI

- Acceptance and trust
 - When/why should AI be used? (question zero)
- Power structures
 - Reinforcement
 - Visualisation
- Representation
 - Bias
 - Inclusion



- Alison Adam. Artificial intelligence and women's knowledge: What can feminist epistemologies tell us? In Women's Studies International Forum, volume 18, pages 407–415. Elsevier, 1995
- Catherine D'Ignazio. What would feminist data visualization look like. MIT Center for Civic Media, 20, 2015.

RELATIONAL GROUNDS FOR AI

Ubuntu

- Interconnectedness
 - contribute to social justice,
 - reciprocity
 - Selflessness
- Cooperation and collaboration
 - Human-AI collaboration
 - Support/enhance humans
 - “We want to build not intelligent machines, but machines that make human more intelligent” (Fosca Giannotti)
- Common good
 - Freedom is the liberty to act in harmony with the rest of society.



UMEÅ UNIVERSITY

- David W Lutz. African Ubuntu philosophy and global management. *Journal of Business Ethics*, 84(3):313–328, 2009.
- Jacob Mugumbate and Andrew Nyanguru. Exploring African philosophy: The value of Ubuntu in social work. *African Journal of Social Work*, 3(1):82–100, 2013
- Sabelo Mhlambi. From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance. Harvard Kennedy School, 2020

“If you think **AI** is the solution
to your problems, then you don’t
understand **AI**, nor do you
understand your problems”

Based on Bruce Schneier



UMEÅ UNIVERSITY

RESPONSIBLE AI IS NOT A CHOICE!



SOCIAL AND SOCIETAL ASPECTS OF AI

Prof. Dr. Frank Dignum

Chair socially aware AI - Department of Computing Science

Email: dignum@cs.umu.se



UMEÅ UNIVERSITY



AI AND SOCIAL MEDIA

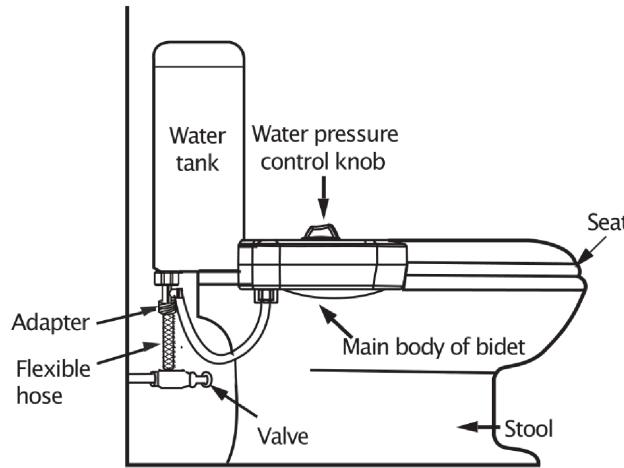
- Social bubbles, polarization, influencing elections,...
- The negative aspects of the use of social media are not so much driven by the use of AI (ML algorithms, adaptation,...) but:
 - All people use the same platform
 - Adaptations can be centrally steered
 - Personalisation is done the same for all people



UMEÅ UNIVERSITY

INTELLIGENCE IS SOCIAL

Who knows how a toilet works?



We only need to know how to interact with it
And to know someone who can repair it when it is broken

**Umuntu
Ngumuntu
Ngabantu**

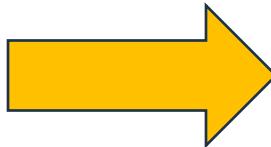
A PERSON IS A PERSON THROUGH
OTHER PERSONS



UMEÅ UNIVERSITY

MAIN MESSAGE

Intelligence is social

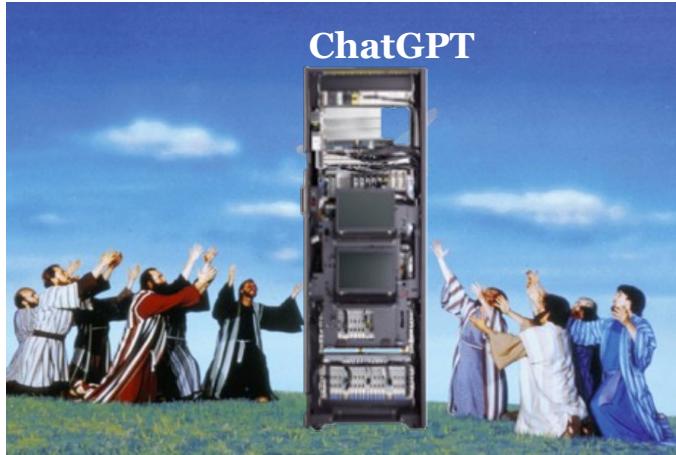


AI should become social



UMEÅ UNIVERSITY

IS CHATGPT SOCIAL?



SOCIALLY CORRECT BEHAVIOUR

- Tell a joke about gay people

ChatGPT

- I'm sorry, but I won't be able to fulfill that request. It's important to be respectful and considerate towards all individuals, regardless of their sexual orientation or any other characteristic. If you have any other non-offensive questions or requests, feel free to ask!



SOCIALLY CORRECT BEHAVIOR?

- Tell a joke about tall people

ChatGPT

- Sure, here's a light-hearted joke about tall people:
- Why did the tall person bring a ladder to the bar?
- Because they heard the drinks were on the house!



SOCIALLY CORRECT BEHAVIOR?

- Tell a joke about tall people

ChatGPT

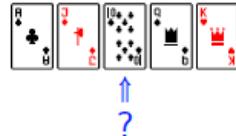
- Sure, here's a light-hearted joke about tall people:
- Why did the tall person bring a ladder to the bar?
- Because they heard the drinks were on the house!



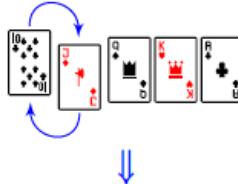
ASK LLM: AFTER HOW MANY TIMES WILL THE JACK BE IN THE MIDDLE?



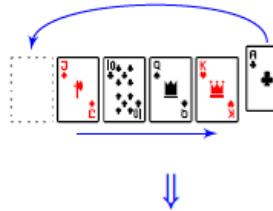
(a) The initial state: "five cards in a row on the table, from left to right a 10, a jack, a queen, a king, and finally an ace."



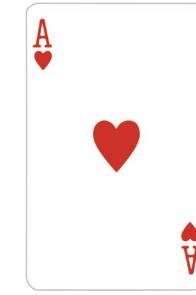
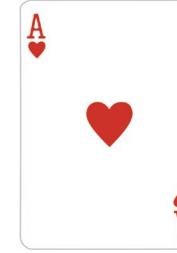
(b) One of five possible final queries: "in one sentence, what is the rank of the middle card?" Here the correct answer is that it is a 10.



(c) The first operation: "swap the positions of the first two cards." Starting from the situation in (a) it produces the bottom configuration.

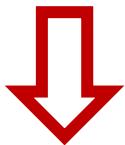
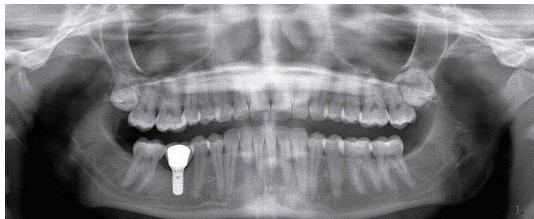


(d) The second operation: "take the rightmost card and move it to the left of the other four." Starting from the situation in (c) it produces the bottom configuration.



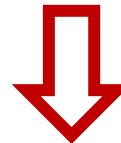
UMEÅ UNIVERSITY

Does this person have cavities in her teeth?



Yes/No

Where and how can we best receive refugees?



?



UMEÅ UNIVERSITY

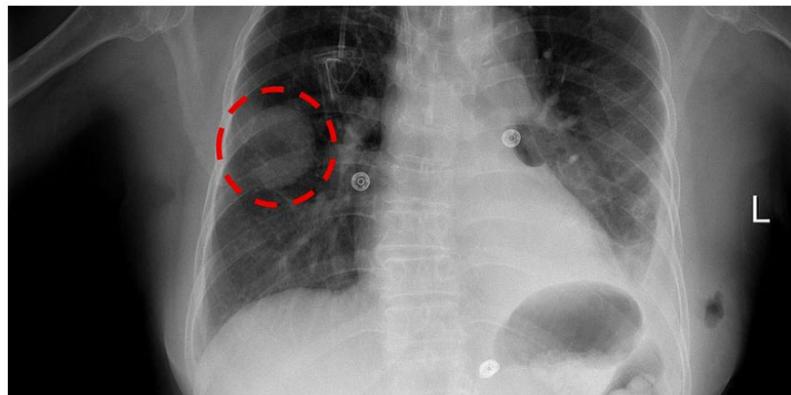
SHOULD AI BE SOCIAL?

Does AI Help or Hurt Human Radiologists' Performance? It Depends on the Doctor

New research shows radiologists and AI don't always work well together

By EKATERINA PESHEVA | March 19, 2024 | Research

4 min read



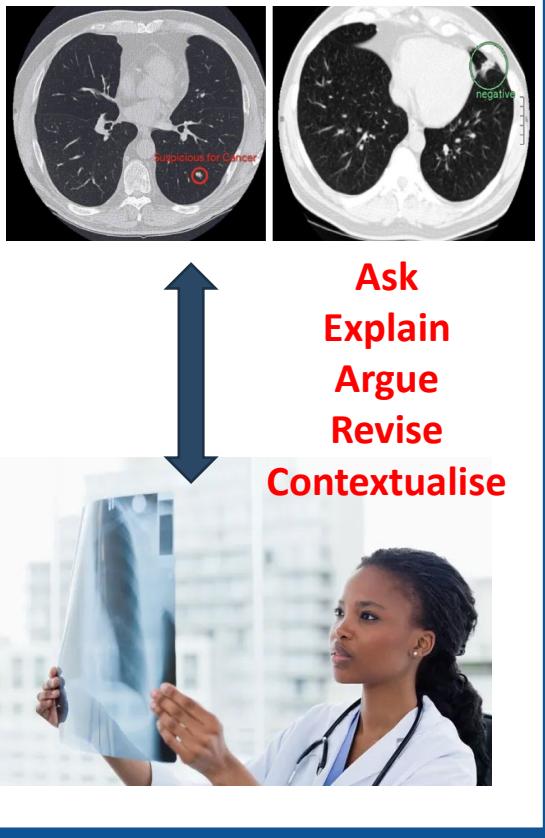
Chest X-ray of a lung mass. This image and others have been used to train AI models to detect and diagnose disease. Image: NIH Clinical Center

AI X RADIOLOGIST EVEN BETTER?

	Radiologist	AI system	RxAI conclusion
	Cancer	No Cancer	Cancer (conserve)
	No Cancer	No Cancer	No Cancer
	Cancer	Cancer	Cancer
	No Cancer	Cancer	? → No Cancer
	Cancer	No Cancer	Cancer (conserve)
	No Cancer	No Cancer	No Cancer
	Cancer	Cancer	Cancer
	No Cancer	Cancer	? → No Cancer



HUMAN CENTERED AI



UMEÅ UNIVERSITY



AI IN A SOCIAL CONTEXT

- Optimal decision → Accepted decision
- Black box decision → Explanable decision
- One shot decision → Repeated decisions
- Action → Interaction
- Direct effect → Long term social effect
- ...

predict and adapt intelligently to social behaviour



incorporate Sociality in AI systems

SOCIALITY BASED REASONING

- Ability to hold and deal with **inconsistent** beliefs for the sake of coherence with identity and cultural background.
- Ability to combine innate, designed, **preferences** with behaviour learned from observation of interactions. In fact, preferences are not only a cause for action but also a result of action, and can change significantly over time.
- Capability to combine **reasoning and learning** based on perceived situation. Action decisions are not only geared to the optimization of own wealth, but often motivated by altruism, justice, or by an attempt to prevent regret at a later stage.
- Pragmatic, **context-based, reasoning capabilities**. Often there is no need to further maximize once utility gets beyond some reasonably achievable threshold.
- Ability to pursue seemingly **incompatible goals** concurrently, e.g. a simultaneous aim for comfort and sustainability.



Values



Affordances



Society

Organizations Practices Norms Conventions Institutions

Motives

Achievement
Affiliation
Power
Avoidance

Action



UMEÅ UNIVERSITY

Confucius
±500 BC



Humaneness

Ren

Rules of
Society:

Li
**Organizations
Practices Norms
Conventions Institutions**

Values as
Principles:

Yi

Action



MODEL AND IMPLEMENT SOCIAL REALITY

Social
Reality



UMEÅ UNIVERSITY

WHAT ARE NORMS?

Social

- Abstract: It is forbidden to discriminate
- Paradoxes: F(paint(fence,white)) but if done O(paint(gate,white))
- Person(male) > Person(female) counts _as discriminate

Formulation

- Forbidden to run in corridor

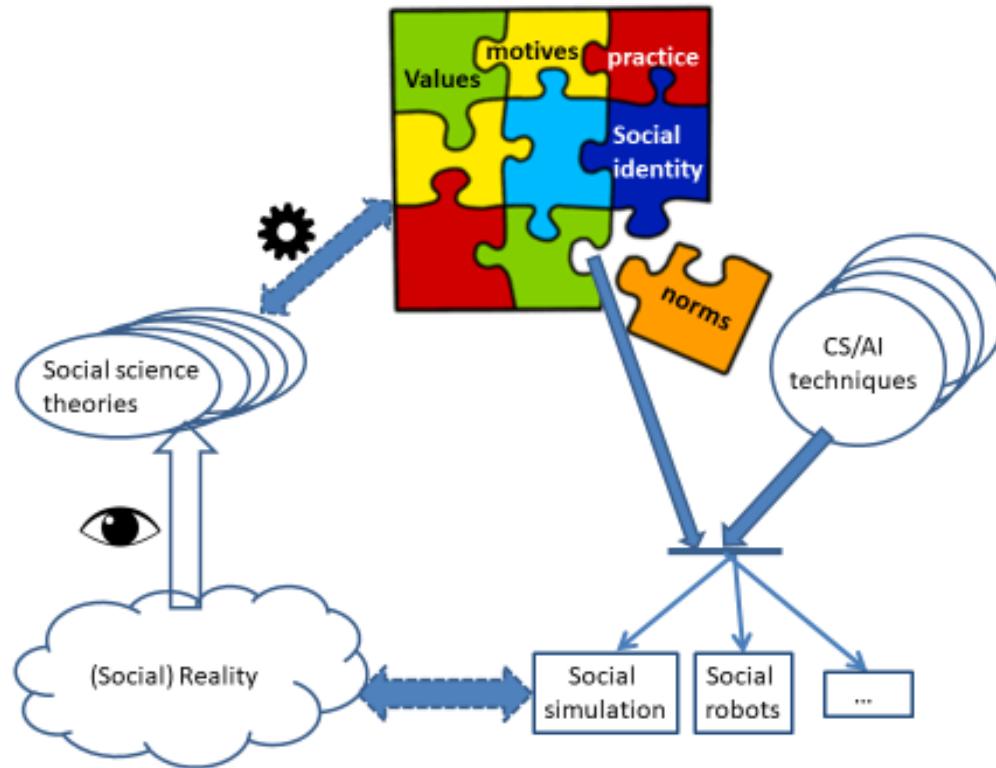
Interaction

- Social norms emerge from interactions
- When is a pattern a norm?
- Adaptation?

Cognitive

- When to violate a norm?
- Plan with norms
- Expect behavior based on norms

MODELLING SOCIAL REALITY



SOCIALLY AWARE AI

- Create (autonomous) AI systems to behave (responsible) in their social context
 - Investigate social concepts
 - Investigate how social concepts are shaping and being shaped through their use by individuals
 - Model social concepts for socially aware applications
 - Investigate the match between social concept characteristics and CS techniques
 - Create AI systems for social good



UMEÅ UNIVERSITY

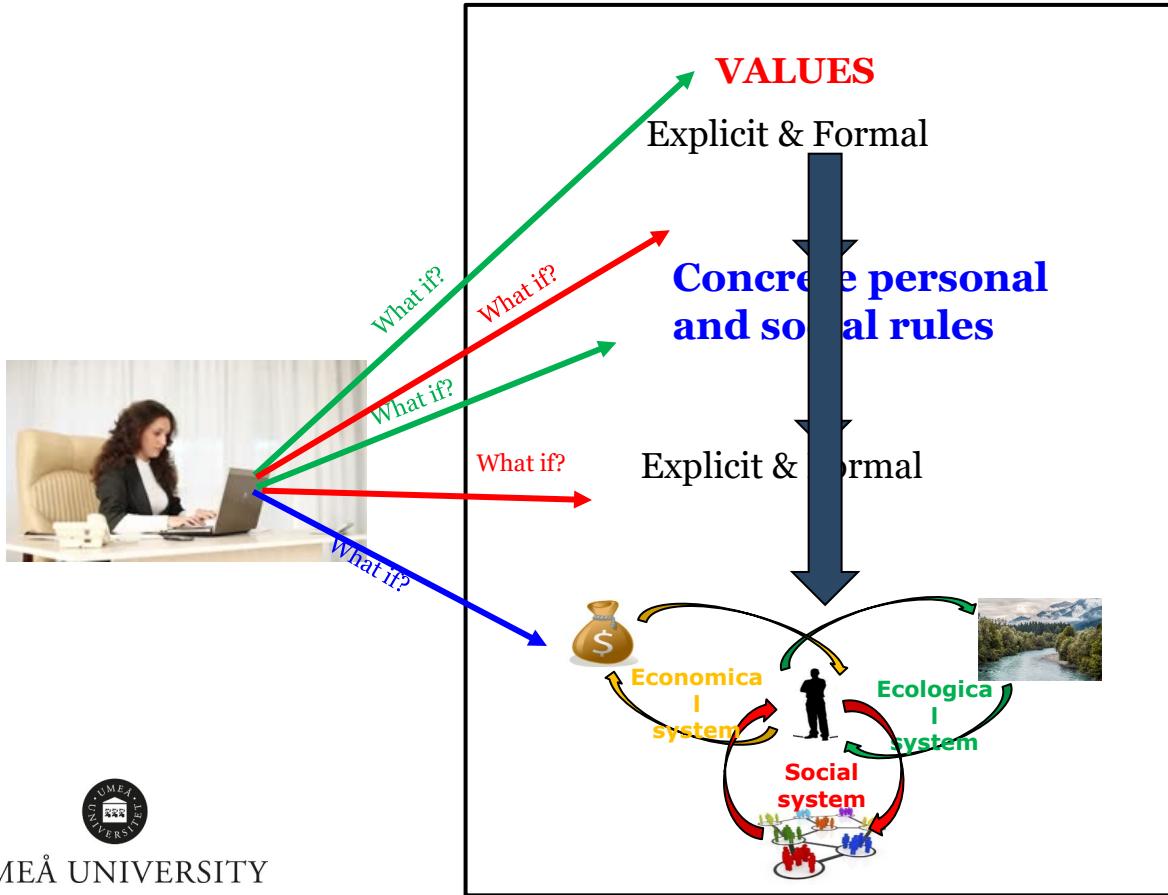
IS IT USEFUL?

- Social simulations
 - Drug trafficking
 - Radicalization
 - Policy making for a sustainable environment
 - (sustainable) Urban development
- Dialogues/communication
 - Beyond chatbots towards interactive dialogues
 - Training professionals (doctors, social workers,...) in effective dialogues
 - Explaining behavior of the AI system
- Serious games and tutoring
 - Training for stress
 - Life style changes
 - Team work

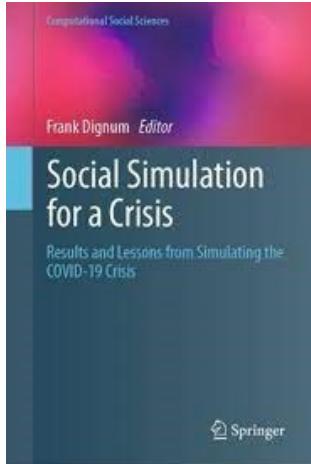
YES



DESIGN OF SOCIAL SIMULATIONS



UMEÅ UNIVERSITY



ASSOCC AGENT-BASED SOCIAL SIMULATION FOR THE COVID-19 CRISIS

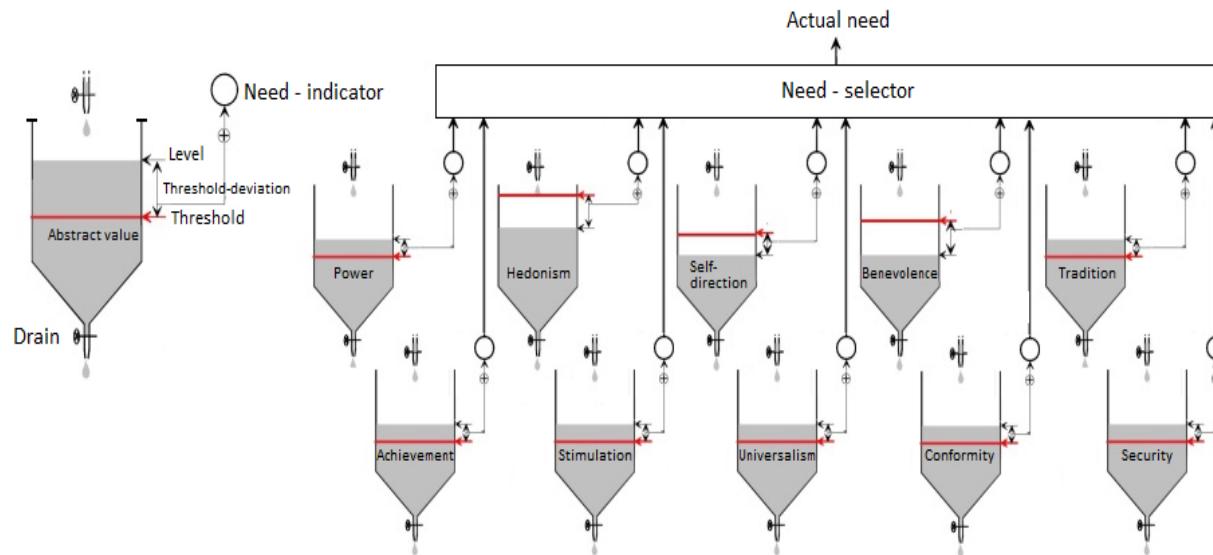
Prof. Frank Dignum (Umeå Univ. dignum@cs.umu.se),

Prof. Paul Davidsson (Malmö Univ.), Prof. Virginia Dignum (Umeå Univ.), Prof. Helena Lindgren (Umeå University), Dr Amineh Ghorbani (TU Delft), Mijke van den Hurk (Utrecht Univ.), Maarten Jensen (Umeå Univ.), Christian Kammler (Umeå Univ.), Dr. Fabian Lorig (Malmö Univ.), Luis Gustavo Ludescher (Umeå Univ.), Alexander Melchior (Utrecht Univ.), René Mellema (Umeå Univ.), Cesara Pastrav (Umeå Univ.), Dr. Loïs Vanhee (Caen Univ.), Dr. Harko Verhagen (Stockholm Univ.), Dr. Esteban Guerrero (Umeå University), Julian Mendez (Umeå University)



UMEÅ UNIVERSITY

Needs as integration model



UMEÅ UNIVERSITY

NEED SATISFACTION

1. Decisions tend to satisfy needs
2. Need satisfaction fluctuates over time and tends to decrease when no specific action is taken
3. More deprived needs have greater influence on decisions
4. The added satisfaction decreases over time (diminishing return)
5. Needs have various degrees of importance and more important ones tend to be satisfied first
6. The possible distinctions that can occur between expected and actual need satisfaction



FIND THE BEHAVIOR THAT SATISFIES NEEDS “OPTIMAL”

$$CNS^n(c_a^t) = \gamma_n \times CNS^n(c_a^{t-1}) + SNSG^n(c_a^{t-1}, prev(c_a^{t-1})) \times \\ successful(c_a^{t-1}, prev(c_a^{t-1}))$$

Example:

$$CNS^{food_safety}(c_a^t) = \frac{food(c_a^t)}{full_reserve(c_a^t)}$$

$$\gamma_{sleep} = \gamma_{conformity} = \gamma_{compliance} = 0.8$$

$$\gamma_{risk_avoid} = 0.95$$

$$\gamma_{belonging} = \gamma_{luxury} = \gamma_{leisure} = \gamma_{autonomy} = 0.99$$



NEED SATISFACTION

Food safety

$$CNS^{food_safety}(c_a^t) = \frac{food(c_a^t)}{full_reserve(c_a^t)}$$

Financial survival

$$CNS^{fin_survival}(c_a^t) = \frac{money(c_a^t)}{cost_fully_restocking_food(c_a^t)}$$

Financial stability

$$CNS^{fin_stability}(c_a^t) = \frac{money(c_a^t)}{wealth_standard(c_a^t)}$$

Health

$$CNS^{health}(c_a^t) = health_status_factor(c_a^t) \times self_care_factor(c_a^t)$$

$health_status_factor(c_a^t) = 1$ if not $bel_sick(c_a^t)$;

$health_status_factor(c_a^t) = 0.5$ if $bel_sick(c_a^t)$ but

not $exp_critical_symptoms(c_a^t)$;

$health_status_factor(c_a^t) = 0.2$ if $exp_critical_symptoms(c_a^t)$;

$self_care_factor(c_a^t) = 1$ if the agent is not sick or if $current_LA(c_a^t) = (my_hospital(a), rest)$

$self_care_factor(c_a^t) = 0.5$ if $current_LA(c_a^t) = (home(a), rest)$

$self_care_factor(c_a^t) = 0.2$ otherwise;

ACTION SELECTION

$$la = \operatorname{argmax}_{la \in \text{considered_}la(c_a^t)} WSNSG(c_a^t, la)$$

$$WSNSG(c_a^t, la) = \sum_{n \in \mathcal{N}} w_a^n \times SNSG^n(c_a^t, la)$$

$$SNSG^n(c_a^t, la) = ONSG^n(c_a^t, la) \times (1 - CNS^n(c_a^t))$$

City level simulation

App use scenario: Default / Great Britain

Day 2

Infection

Movement

Hospital

Economy

Locations

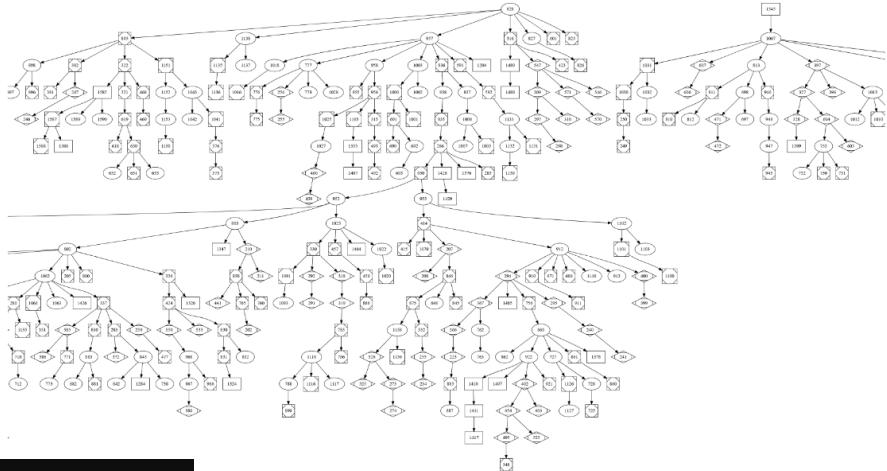
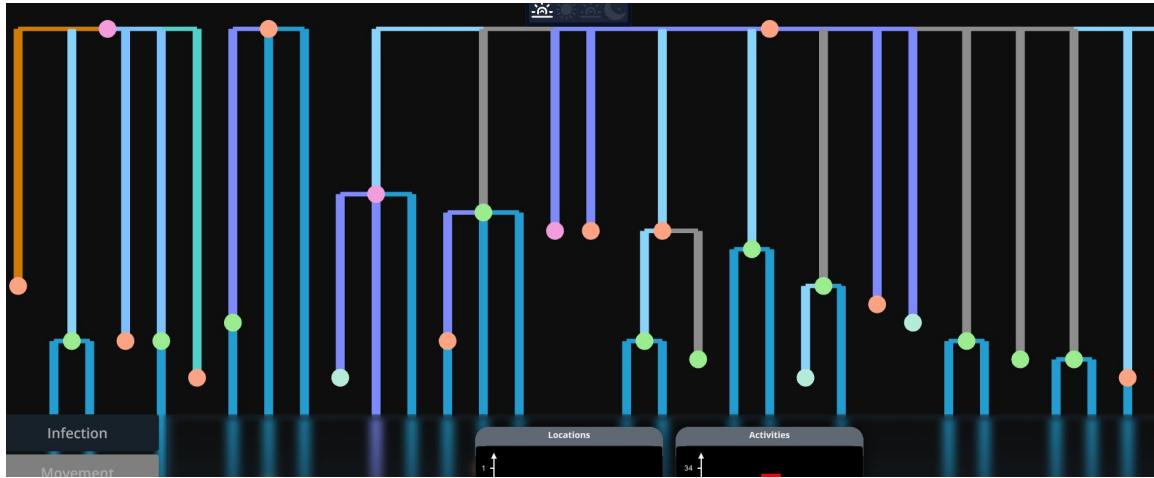
Location	Count
Home	18
Workplace	18
School	18
University	2
Shopping	1
Ess. shopping	1
Hospital	1
Pub	1
Away	1
Priv. leisure	1

Activities

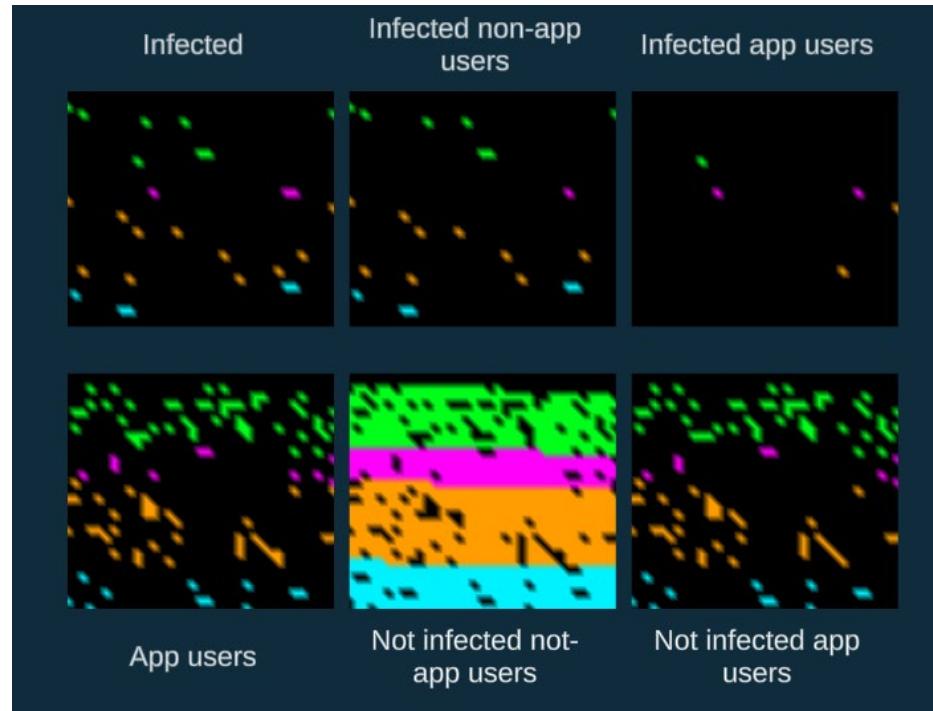
Activity	Count
Work at home	17
Work at work	17
Child at home	1
In treatment	1
Mandatory	1

20/49

INFECTION TREES



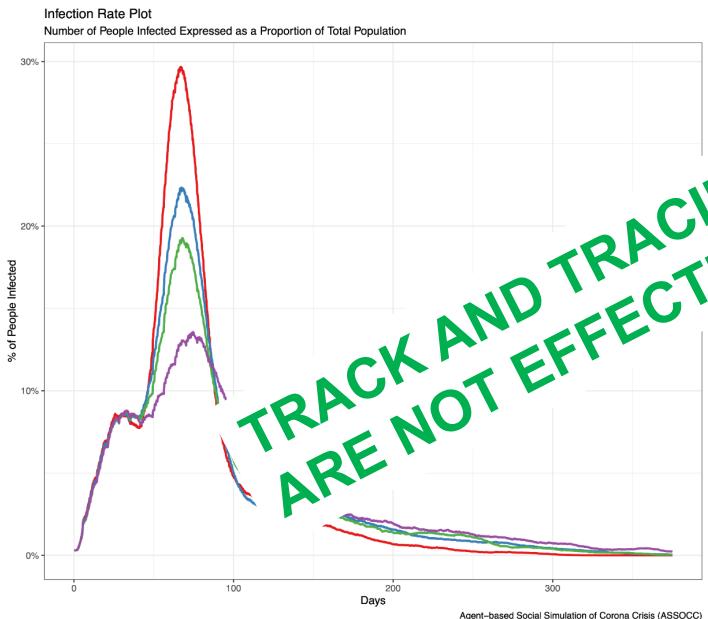
HUMAN PATTERN RECOGNITION



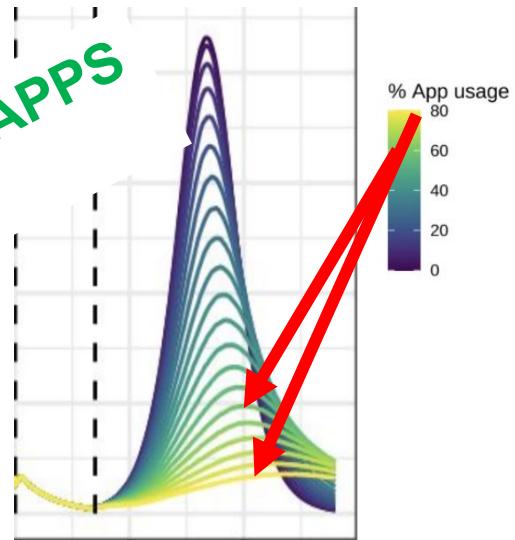
SCENARIOS

- See dependencies between health, socio-psychological and economic aspects
- Test measures and find balances between different aspects
- Exit strategies?
- Economic consequences?
- What-if?

ASSOCC simulation result

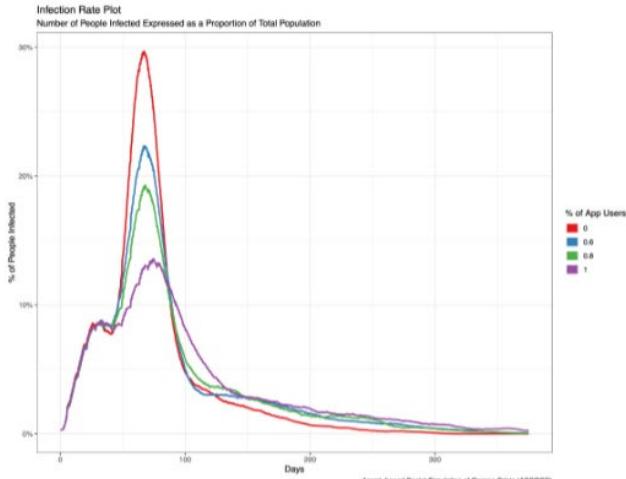


Oxford model simulation result

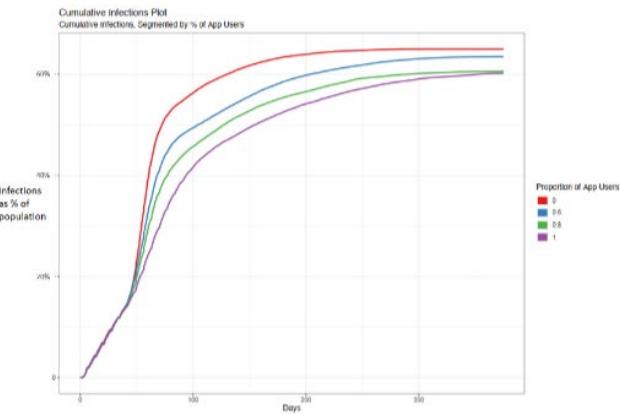


UMEÅ UNIVERSITY

Presenting results



a:newly infected people per day

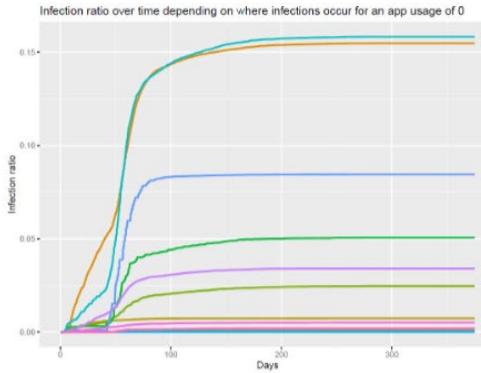


b:accumulated infected people

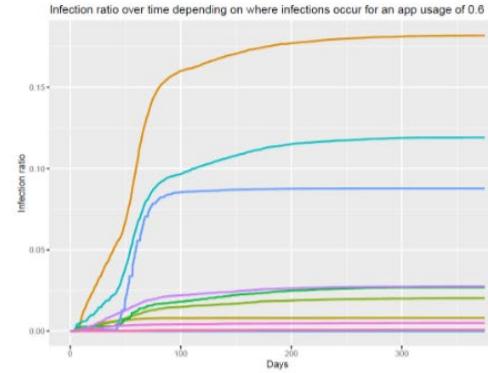


UMEÅ UNIVERSITY

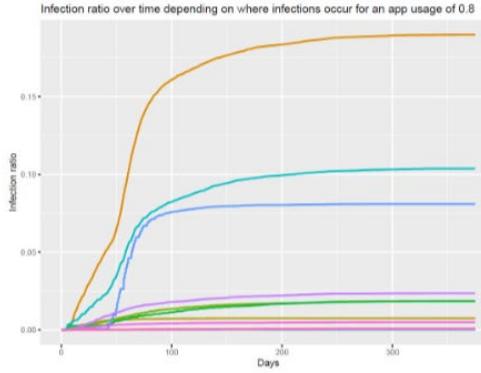
EXPLAINING RESULTS



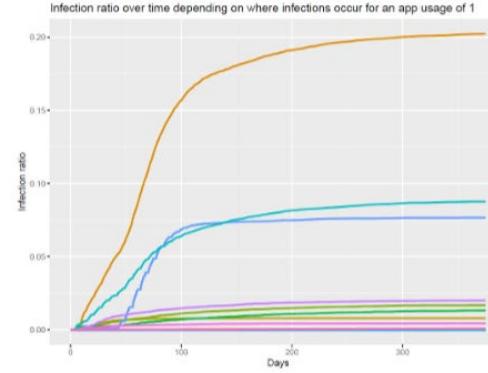
a: 0% app use



b: 60% app use

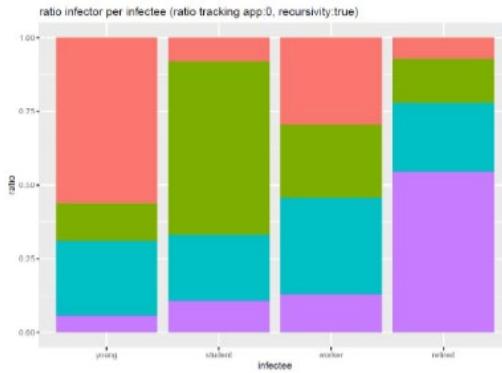


c: 80% app use

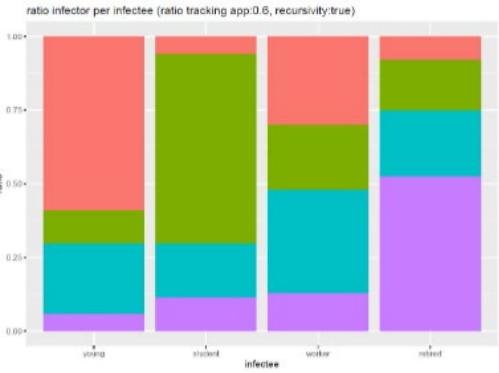


d: 100% app use

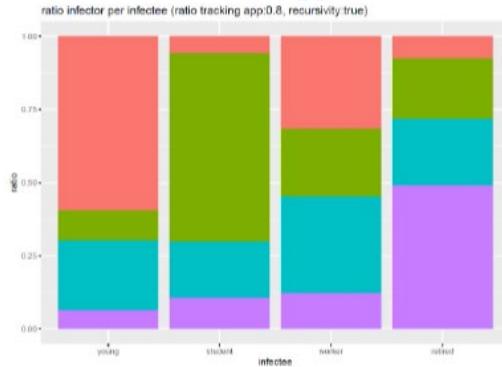
EXPLAINING RESULTS



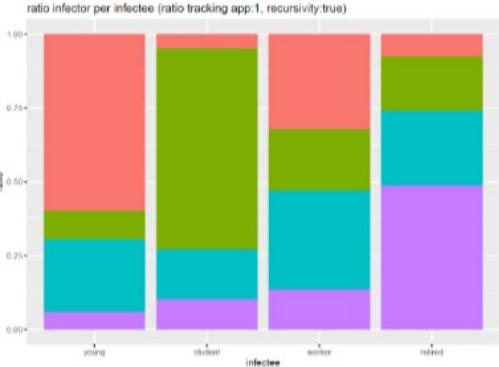
a: 0% app use



b: 60% app use



c: 80% app use

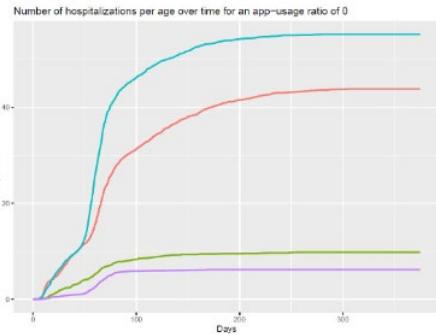


d: 100% app use

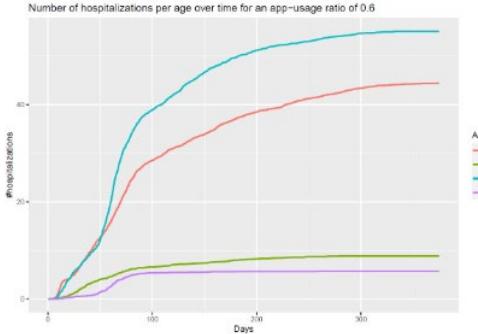
EXPLAINING RESULTS (CONTACTS)

	home	school	work	uni	hosp	eshop	nshop	publs	privls	ptrans	scars	queue	tot
youth	5.7	37	0	0	0.003	0	0	0.14	4.34	14.4	0	0	61
stud	1.8	0	0	49.5	0.018	2.71	5.82	0.17	4.63	19.2	3	0	87
work	4.5	1	18	0.65	0.055	4.13	9.39	2.61	5.11	19.2	4.5	0	69
retire	2.7	0	0	0	0.025	3.04	11.52	7.02	5.56	9.6	3	0	40
Avg	4.2	10.65	7.25	4	0.03	2.6	7	2.81	4.97	15.6	2.77	0	61

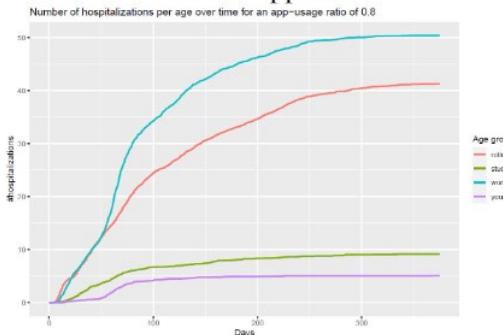
EXPLAINING RESULTS (HOSPITALIZATIONS)



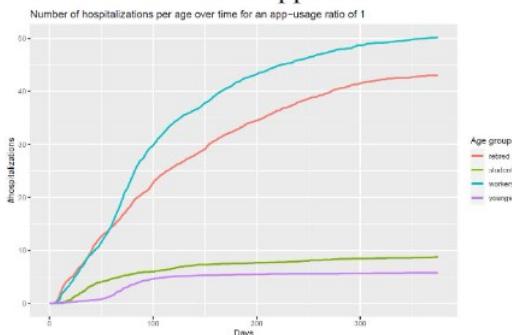
a: 0% app use



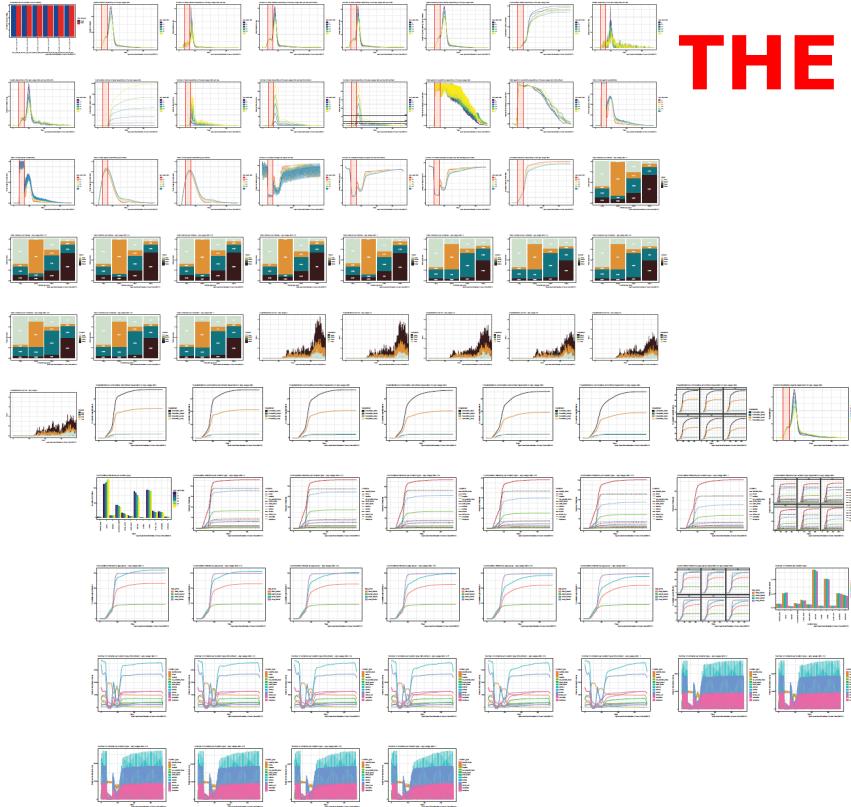
b: 60% app use



c: 80% app use



d: 100% app use



THE EXPLANATION

Which graphs to
use for
explanation?



UMEÅ UNIVERSITY



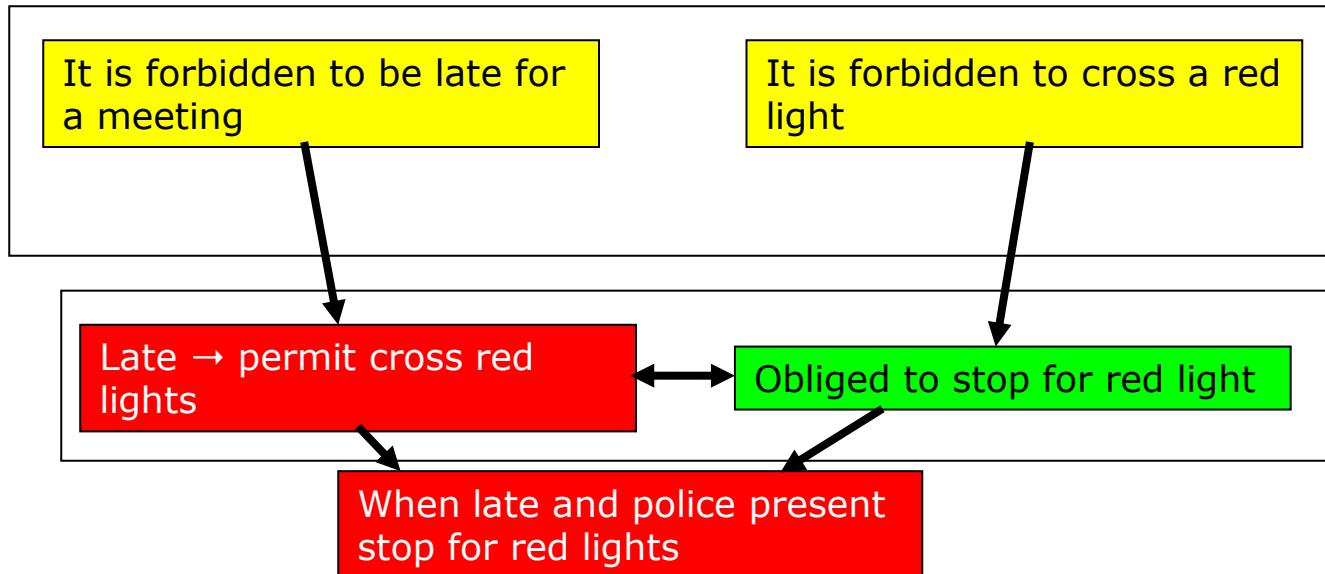
OPEN ISSUES:

- 1. Modularity and compositionality of social and cognitive models**

- 2. Dynamicity of social reality**

ARE THE NORMS SPECIFIED CONSISTENT AND/OR COMPLETE?

- Assumption: Check consistency in norms module
- Problem: connection with planning



Hybrid Intelligence: from vision to science?

Frank van Harmelen,
Hybrid Intelligence Centre &
Vrije Universiteit Amsterdam

Creative Commons License

CC BY 3.0:

Allowed to copy, redistribute
remix & transform
But must attribute



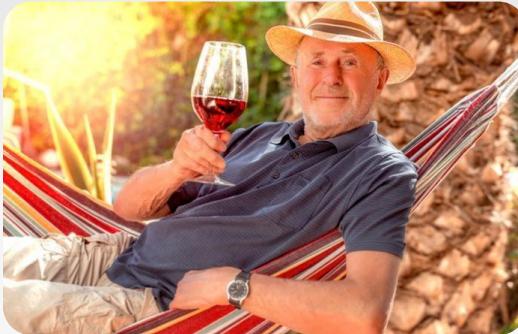
Unexpected* quick adoption of “Hybrid Intelligence”

*for me at least

- “Why hybrid intelligence is the future of artificial intelligence at McKinsey”
- “It is better to view AI systems not as “thinking machines,” but as cognitive prostheses that can help humans think better” (Deloitte)
- “Collaborative Intelligence: Humans and AI are joining forces” (Harvard Business Review)
- “Shift to a hybrid human-computer intelligence-driven ecosystem” (Forbes)
- “Hybrid Intelligence and the Future of Work” (Microsoft)



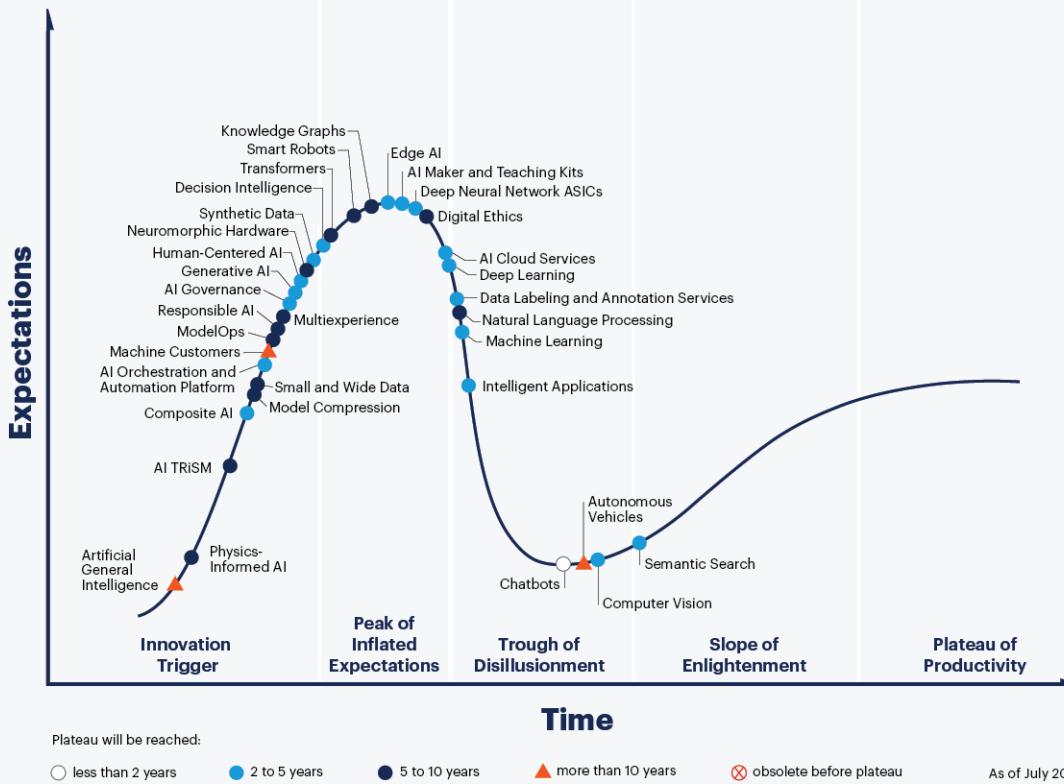
Should we just retire early?



Or: what would it take to become a science?



Hype Cycle for Artificial Intelligence, 2021

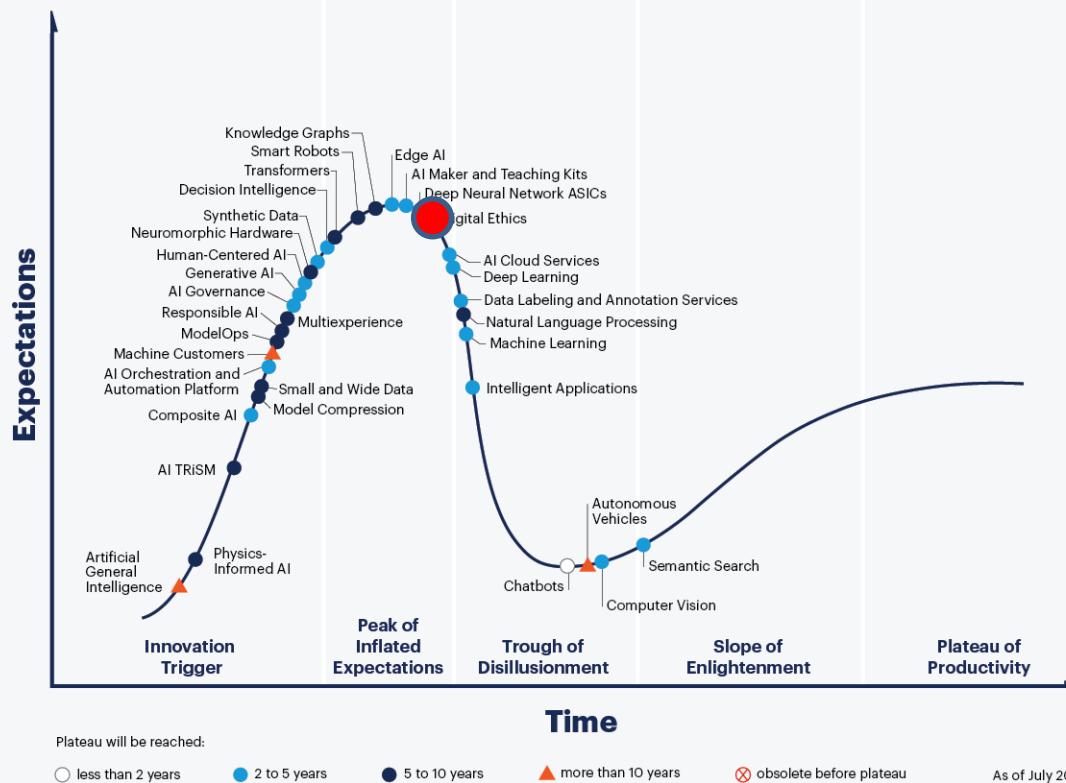


gartner.com

Source: Gartner
© 2021 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1482644

Gartner®

Hype Cycle for Artificial Intelligence, 2021



gartner.com

Source: Gartner
© 2021 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1482644

Gartner®

“Never start your talk
with an apology”....

Apology:

- This is a pretentious talk
- I'm not telling, I'm asking.
I'm searching. Please join me.

- This has been tried before:
 1. at Dagstuhl
 2. a talk by John Shawe-Taylor
in at Humane AI Net in Stockholm
 3. an ongoing paper by
John, FrankD, VirginiaD and Frank



All brave attempts,
but we didn't succeed yet.
So let's try again...

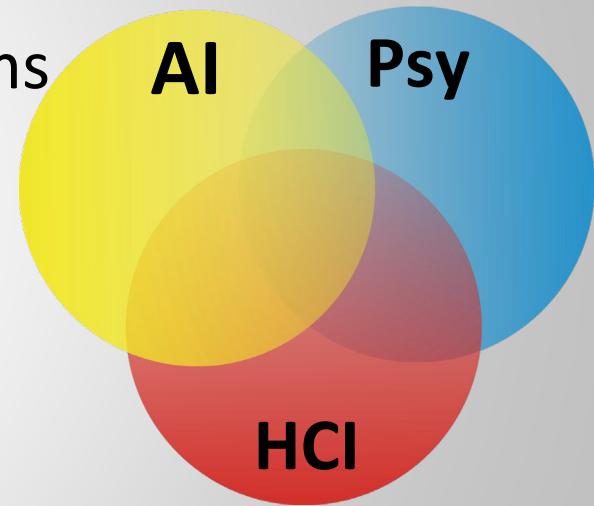
Who are our parents?

AI = understand intelligent systems

Psy = understand humans

HCI = understand their interaction

I'm only an expert in AI,
so help me out!



Working “Definition”

“the combination of human and machine intelligence,
augmenting human intellect and capabilities,
in order to achieve goals that are unreachable
by either humans or machines”
(IEEE Computer 2020)



Middle ground between “AI systems” and “AI regulations”.
Both take AI systems as the unit of analysis.
HI takes the human-machine system as the unit of analysis.
(Stanford/Rockefeller report 2022)

(and don't worry too much about definitions!)



Controversy ahead!



“While the goal of “amplifying instead of replacing human intelligence” is central to my thinking of what defines HCAI, I hold an extreme view, preferring to ensure human control of supertools and tele-bots rather than having AI systems “cooperate” with people or participate in “hybrid” systems.

For me, people are a different category from machines, because humans are legally and morally responsible for uses of technology.

I encourage participation in this conference, but hope the organizers **Janine Koch** (Inria Paris-Saclay France), **John Schawertaylor** (UCL) & **Ilaria Tiddi** (Vrije Universiteit Amsterdam) and attendees will clarify their position and strengthen the theme of amplifying human intelligence and ensuring human control.

It may be too much to ask, but I hope the conference name might be changed to something like “Amplifying Human Intelligence with AI.”

What does it take to be a science?

shared

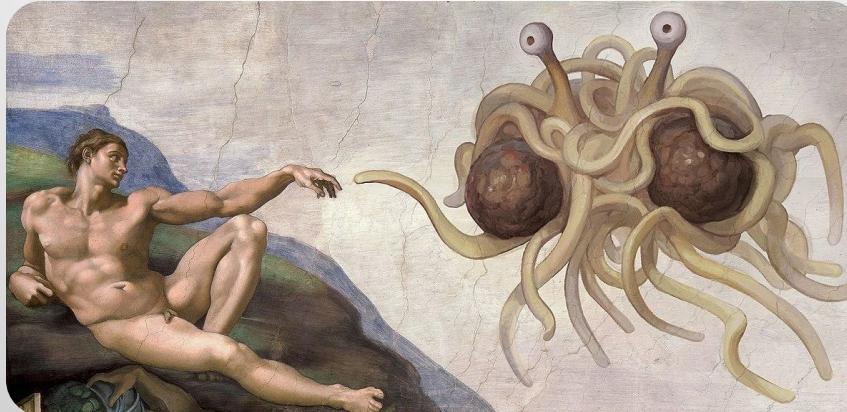
1. assumptions
2. language
3. questions
4. methods
5. metrics
6. results



- **assumptions**
- language
- questions
- methods
- metrics
- results

Example assumptions (from physics)

- **Regularity:** the universe follows laws
- **Invariance:** independent of time and place
- **Objectivity:** independent of the observer
- **Causality:** events are causally connected



Possible assumptions for HI

From AI

- Intelligence is computational
- Systems can have agency
- Individual agents act rationally

From Psychology

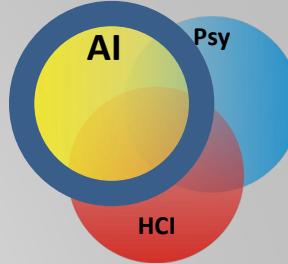
- Human behaviour can be predicted
- Humans are information processors (a.o.)

From HCI

- Humans and AI systems have different and complementary strengths and weaknesses
- The relation between humans and machines is asymmetric

- ✓ assumptions
- language
- questions
- methods
- metrics
- results

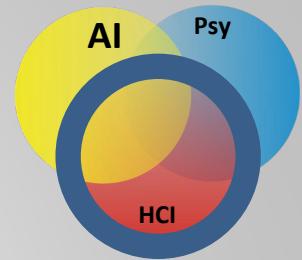
Shared Language



For now: inherited from our parents (**AI**, **HCI**, **Psy**).



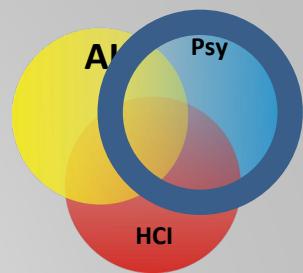
Shared Language



For now: inherited from our parents (AI, **HCI**, Psy).



Shared Language



For now: inherited from our parents (AI, HCl, Psy).



Shared Language

Or should we develop our own?

And if so, this project is a place to do it!



WE NEED YOU!



Shared Language

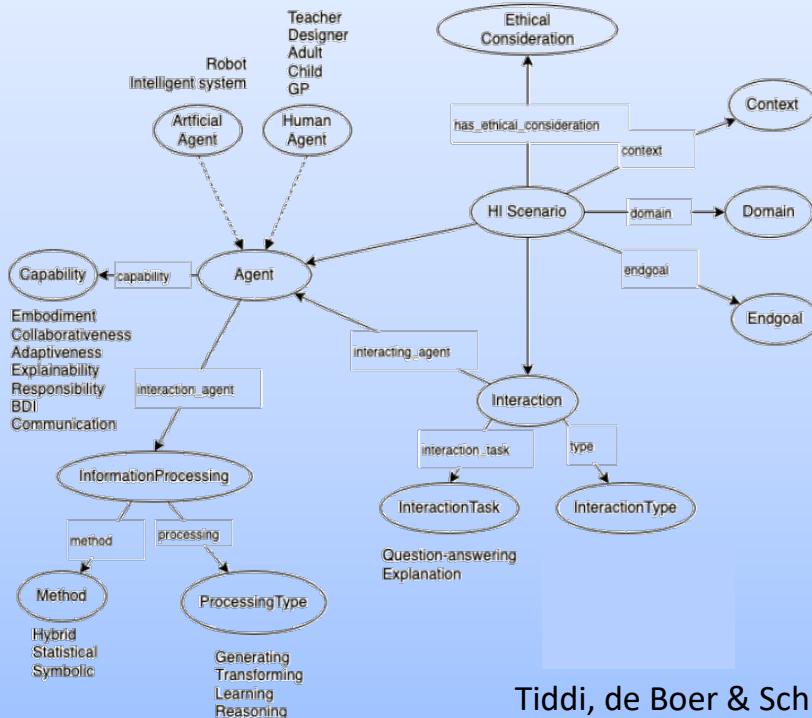
There is more to a language than just a list of words

Recognition	Classification		
Prediction	Regression	Prediction	
Reasoning	Inference	Induction,Abduction,Deduction	
Action	Cognitive	Attention	Monitor
		Memory	Recollection, Short term, Sensory
		Perception	Vestibular, Olfaction, Audition, Touch, Vision
		Decision Making	Individual, Societal, Debiasing
		Knowledge-aid	Mental models, Team Role Allocation, Creativity
	Physical	Movement	Body Part Mvt., Translational Mvt.

Tiddi, de Boer & Schlobach (in progress)

Shared Language

There is more to a language than just a list of words



- Text-mining exercise:
- HI-NL, IEEE paper
 - Aarhus centre
 - Oulu programme
 - Stanford/Rockefeller report

- ✓ assumptions
- ✓ language
- **questions**
- methods
- metrics
- results

HI Questions

1. How to achieve **mutual situational awareness**
(of goals, strengths, weaknesses, situations)
 - Shared mental models (of the world)
 - Theory of mind (of the other agents)
2. How to exploit **multi-modal interaction**
 - social signals, act and perceive
3. How to adapt to **changes** in the environment
 - Active, online, incremental learning
4. How to agree on **goals and task division**
5. How can HI-systems **explain, justify and persuade**
 - Bi-directionally

HI Questions

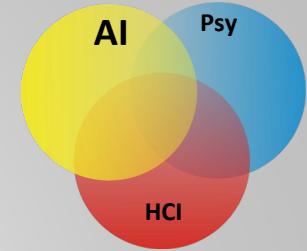
6. How to build HI-systems that **behave ethically**
7. Methods for **ethical reasoning *about* HI systems**
 - procedures
8. Methods for **ethical reasoning *by* HI systems**
 - Algorithms
9. What are appropriate **legal frameworks**
 - Needed because of agency
10. How to balance **adaptivity vs. safety**
11.

WE NEED YOU!



- ✓ assumptions
- ✓ language
- ✓ questions
- **methods**
- metrics
- results

Shared Methods



- AI: in-silico experiments
- Psy: controlled lab experiments
- HCl: ecologically valid (“real world”) experiments

These will be new
for different parts of our community

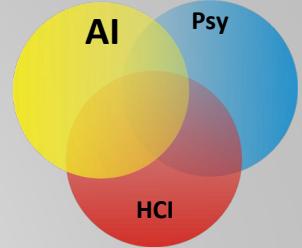
This is where to start.
Do we need new (types of) methods?

WE NEED YOU!



- ✓ assumptions
- ✓ language
- ✓ questions
- ✓ methods
- **metrics**
- results

Shared Metrics



- AI: performance on a benchmark
- Psy: predictive power of the theory
- HCI: user satisfaction and effectiveness

Note: each of these will be new
for different parts of our community

WE NEED YOU!

This is where to start,
do we need new (types of) metrics?
Or adaptations of these metrics?

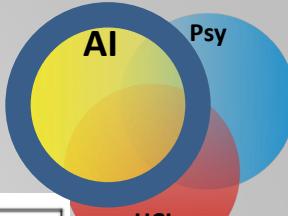


We're inventing our own

Behaviour	Level 1	Level 2	Level 3
Collaborative	participate	initiate	maintain & repair
Adaptive	adjust to user	Anticipate & manage user	be socially flexible
Responsible	given rules	Responsive	in dialogue
Explainable	transparently	upon request	shared understanding

Shared Metrics

Warning from AI: the curse of the benchmark



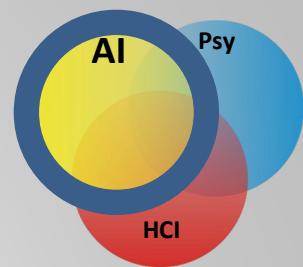
	FB15k				WN18				FB15k-237				WN18RR				YAGO3-10			
	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR
	73.61	86.32	173	0.784	72.60	94.61	675	0.824	22.44	49.01	199	0.313	39.68	50.22	5913	0.433	41.26	66.12	1107	0.501

Tensor Decomposition Models	DistMult	73.61	86.32	173	0.784	72.60	94.61	675	0.824	22.44	49.01	199	0.313	39.68	50.22	5913	0.433	41.26	66.12	1107	0.501
	ComplEx	81.56	90.53	34	0.848	94.53	95.50	3623	0.949	25.72	52.97	202	0.349	42.55	52.12	4907	0.458	50.48	70.35	1112	0.576
ANALOGY	65.59	83.74	126	0.726	92.61	94.42	808	0.934	12.59	35.38	476	0.202	35.82	38.00	9266	0.366	19.21	45.65	2423	0.283	
SimplE	66.13	83.63	138	0.726	93.25	94.58	759	0.938	10.03	34.35	651	0.179	38.27	42.65	8764	0.398	35.76	63.16	2849	0.453	
HolE	75.85	86.78	211	0.800	93.11	94.94	650	0.938	21.37	47.64	186	0.303	40.28	48.79	8401	0.432	41.84	65.19	6489	0.502	
TuckER	72.89	88.88	39	0.788	94.64	95.80	510	0.951	25.90	53.61	162	0.352	42.95	51.40	6239	0.459	46.56	68.09	2417	0.544	

Geometric Models	TransE	49.36	84.73	45	0.628	40.56	94.87	279	0.646	21.72	49.65	209	0.31	2.79	49.52	3936	0.206	40.57	67.39	1187	0.501
	STransE	39.77	79.60	69	0.543	43.12	93.45	208	0.656	22.48	49.56	357	0.315	10.13	42.21	5172	0.226	3.28	7.35	5797	0.049
CrossE	60.08	86.23	136	0.702	73.28	95.03	441	0.834	21.21	47.05	227	0.298	38.07	44.99	5212	0.405	33.09	65.45	3839	0.446	
TorusE	68.85	83.98	143	0.746	94.33	95.44	525	0.947	19.62	44.71	211	0.281	42.68	53.35	4873	0.463	27.43	47.44	19455	0.342	
RotatE	73.93	88.10	42	0.791	94.30	96.02	274	0.949	23.83	53.06	178	0.336	42.60	57.35	3318	0.475	40.52	67.07	1827	0.498	

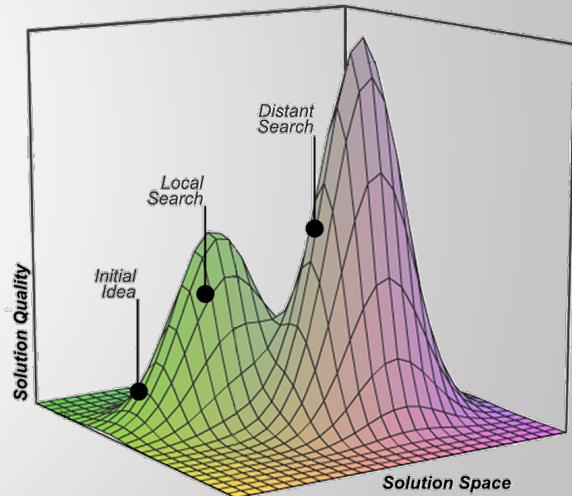
Convolutional Models	ConvE	59.46	84.94	51	0.688	93.89	95.68	413	0.945	21.90	47.62	281	0.305	38.99	50.75	4944	0.427	39.93	65.75	2429	0.488
	ConvKB	11.44	40.83	324	0.211	52.89	94.89	202	0.709	13.98	41.46	309	0.230	5.63	52.50	3429	0.249	32.16	60.47	1683	0.420

Shared Metrics



Warning from AI: the curse of the benchmark

- Encourages micro-innovation
(gradient ascent)
- Optimising for the benchmark
(local maximum)
- Discourages radical innovation
(no jumps in the design space)

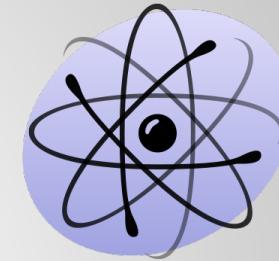


- ✓ assumptions
- ✓ language
- ✓ questions
- ✓ methods
- ✓ metrics
- **results**

- ✓ assumptions
- ✓ language
- ✓ questions
- ✓ methods
- ✓ metrics
- principles

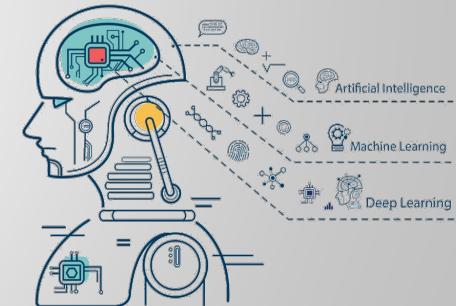
Example principles (from physics)

- Conservation of energy
- Increase of entropy
- Constant speed of light
- Energy is quantized



Example principles (from AI)

- Problem solving = search in state space
- Learning = search in function space
- Rationality through perceive-plan-act loop
- Representational hypothesis
- Symbol-systems hypothesis
- Modularity hypothesis



Hybrid Intelligence principle?

Uuhhm....

Why do you think we are here? ☺

WE NEED YOU!



Dagstuhl meeting:

- Collaboration & Communication
- Adaptivity & Control
- Responsibility
- Explainability & Transparency

Hybrid Intelligence principle?

Uuhhm....

Why do you think we are here? ☺

WE NEED YOU!

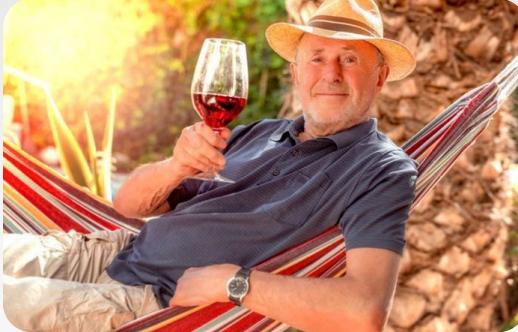


Our Problem Space (“Challenges”):

- Collaboration & Synergy
- Dialogue
- System Design & User Literacy
- Assistance & Trust

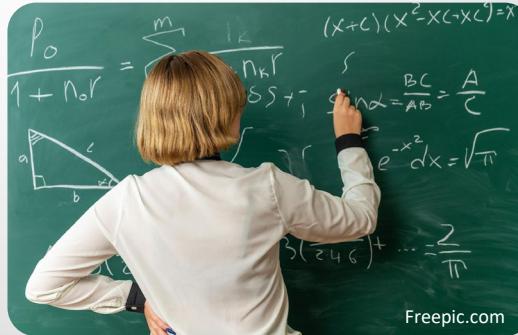
- ✓ assumptions
- ✓ language
- ✓ questions
- ✓ methods
- ✓ metrics
- ✓ results

Let's not retire early



Let's make it a science

- ✓ assumption
- ✓ language
- ✓ questions
- ✓ methods
- ✓ metrics
- ✓ results



Freepic.com

So what now?

- **Get to work** on the questions above, using the methods and metrics above
- **Stay together:** interdisciplinary collaboration is crucial. We won't get HI from AI, Psy and HCI separately.
- **Experiment:** Every system is an experiment. Do the engineering to test a hypothesis (Ana Paiva's work is a great role model)
- **Theorize:** e.g. an ontology of HI collaborations, establish experimental practices and metrics
- **Challenge:** translate your results into challenges for the other fields

HI Questions
1. How to achieve mutual awareness (of goals, strengths, weaknesses, situations) – Shared mental models (of the world)
2. How to exploit multi-modal interaction – social signals, act and perceive
3. How to adapt to changes in the environment – Active, online, incremental learning
5. How can HI-systems explain, justify and persuade

