On the Ethics of Complex Systems

Erich Prem

Why is it so hard to know how to do the right thing in IT and AI?

eu|te|ma
TECHNOLOGY MANAGEMENT

UNIVERSITY OF VIENNA

Some examples of reasons for concern…

- Humans as the subject of machine decision-making
- Loss of state sovereignty
- Control of human behaviour
- Monopolisation of online platforms
- Private and government surveillance
- Fake news, filter bubbles, de-objectified discourse
- Algorithmic decision-making, AI
- Alienation of labour and de-qualification
- Social disconnect

DIGITAL
HUMANISM

What is digital humanism?

DIGITAL HUMANISM

Digital humanism is an initiative to actively shape digitization so that people and society are the focus.

Digital humanism is a call to use digital technologies to protect human rights and develop democracy.

Digital humanism acknowledges the key role of digital technologies for progress and innovation and seeks to expand it to sustain and expand our social achievements.

https://dighum.ec.tuwien.ac.at/dighum-manifesto/

Introducing digital components can have massive changes.

Communication

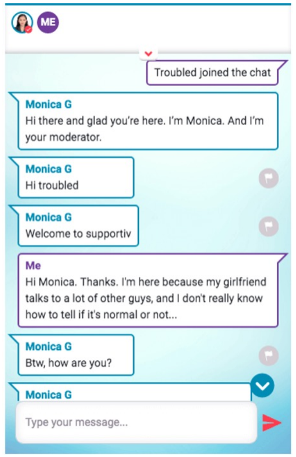
- Monitor
- Change, exclude
- Offer

Car

- Control risk
- Switch-off
- Trace
- …

Electricity

- Monitor
- Switch-off
- Auction
- …



Troubled joined the chat

Monica G
Hi there and glad you're here. I'm Monica. And I'm your moderator.

Monica G
Hi troubled

Monica G
Welcome to supportiv

Me
Hi Monica. Thanks. I'm here because my girlfriend talks to a lot of other guys, and I don't really know how to tell if it's normal or not...

Monica G
Btw, how are you?

Monica G

Previously unobservable behaviour becomes
monitored, accumulated, predicted and controlled.
This may have unwanted, unpredicted, and undesirable consequences.

A chain of digital processes creates new phenomena of surveillance, prediction and control.

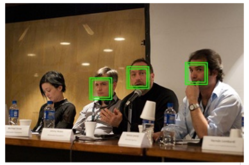
Digital Hardware → Digital service platforms

Online interactions → Harvesting of data

Prediction models → Digital marketing

Digital control → New digital powers

Uber Gmail WienMobil



Was ist guter Content für Google?

Google Ad Manager Google Ads Google Marketing Platform

Philosophy of morality

> *Morality is an informal public system applying to all rational persons, governing behaviour that affects others, and includes what are commonly known as the moral rules, ideals and virtues and has the lessening of evil and harm as its goal.*
> (Bernard Gert)

εθος – custom (behaviour)

ηθος – character (attitude towards behaviours)

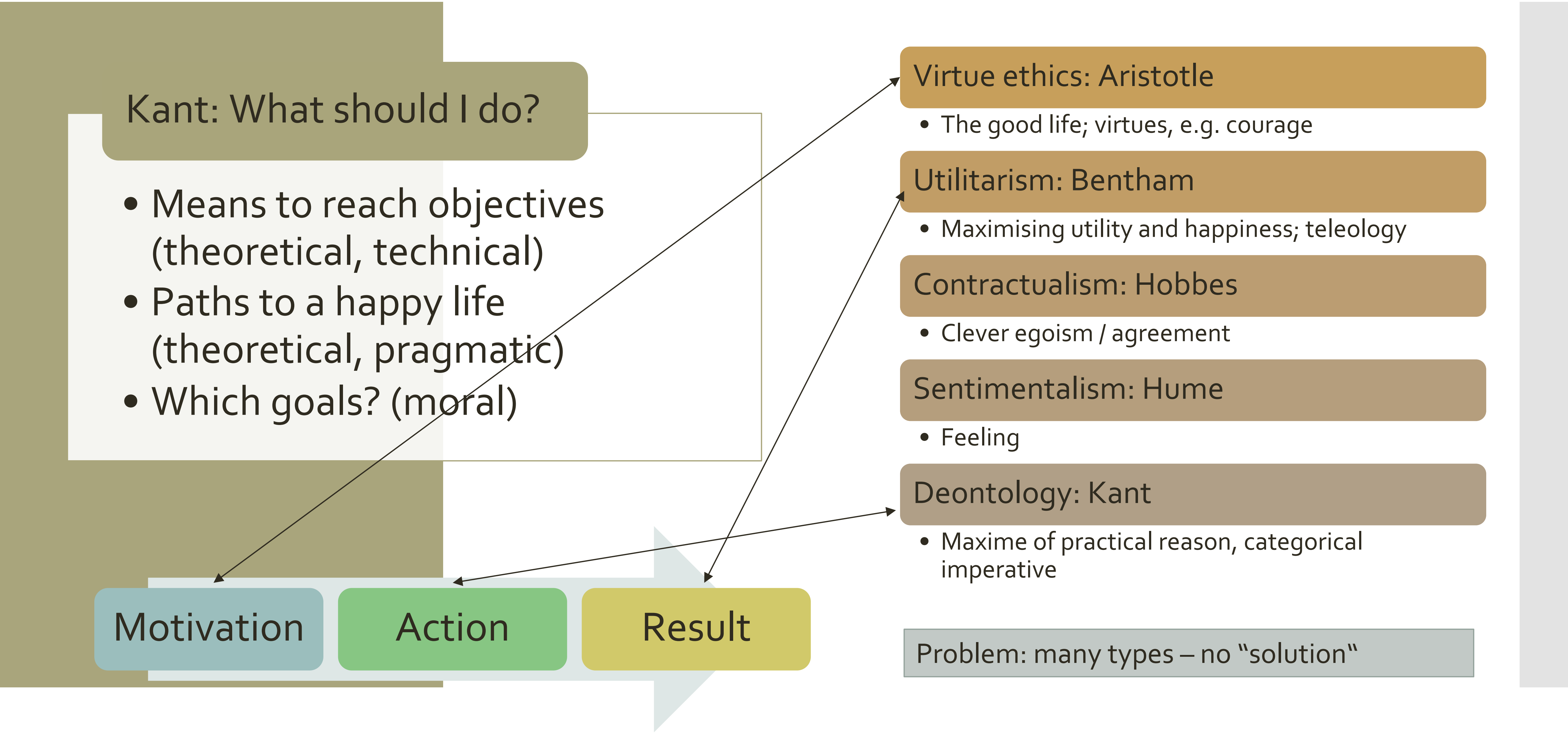
descriptive, normative, applied, metaethics

Some common virtues
- truthfulness
- courage
- honesty
- impartiality
- reliability
- …

Ideals: e.g. justice

Some common harms
- death
- pain
- disability
- loss of freedom
- loss of pleasure
- loss of rights
- …

Types of ethics



Kant: What should I do?

- Means to reach objectives (theoretical, technical)
- Paths to a happy life (theoretical, pragmatic)
- Which goals? (moral)

Virtue ethics: Aristotle

- The good life; virtues, e.g. courage

Utilitarism: Bentham

- Maximising utility and happiness; teleology

Contractualism: Hobbes

- Clever egoism / agreement

Sentimentalism: Hume

- Feeling

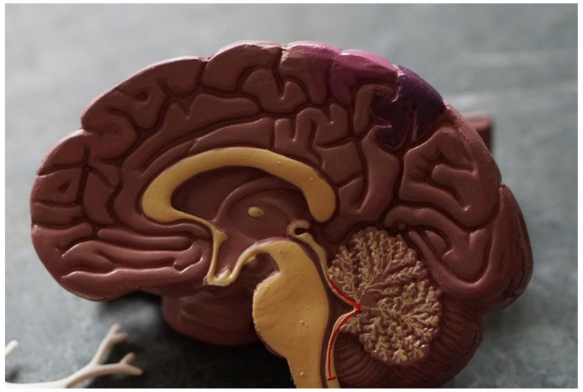
Deontology: Kant

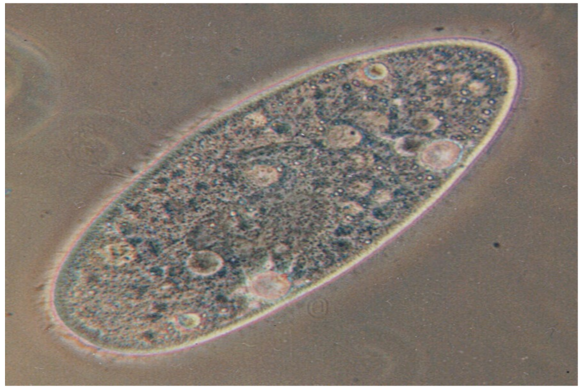
- Maxime of practical reason, categorical imperative

Problem: many types – no “solution“

Complex systems

- Net-like causal structures (high connectivity)
- Nonlinear interactions
- Adaptivity
- Open systems with problematic boundaries
- Choice of observables
- Anticipation and finality

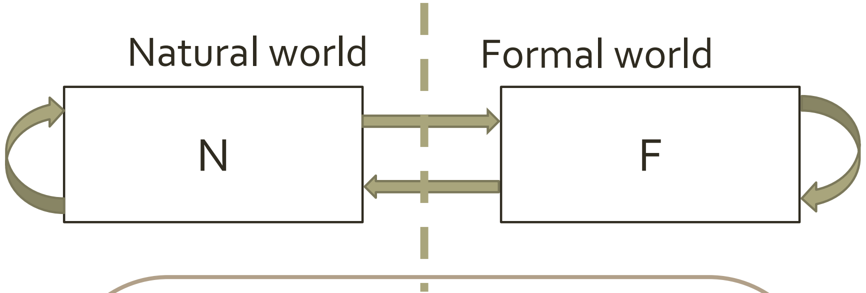




Fotos by Salmen Bejaoui, Robina Weermeijer on Unsplash.
Von Barfooz at the English Wikipedia. - Originally uploaded to the English Wikipedia, where it was made by Barfooz., CC BY-SA 3.0,

The modelling problem

Model
- Correct
- Relevant
- Simple



Natural world

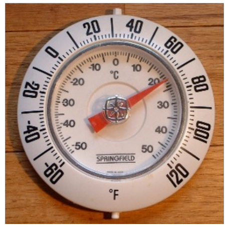
Formal world

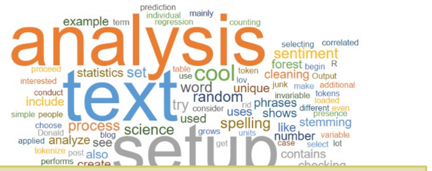
Natural law

N

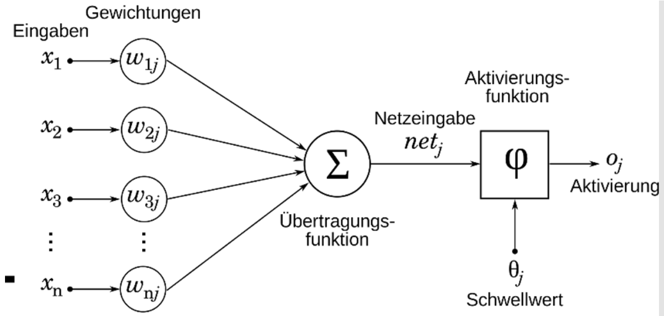
F

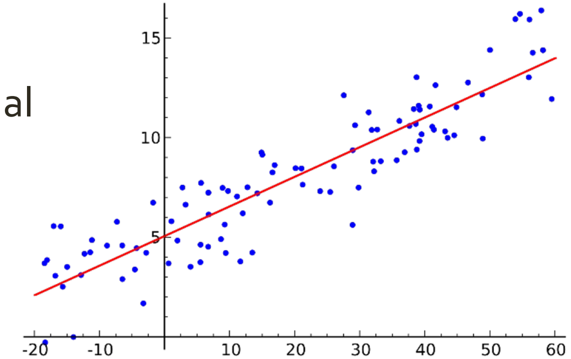
Formal rules

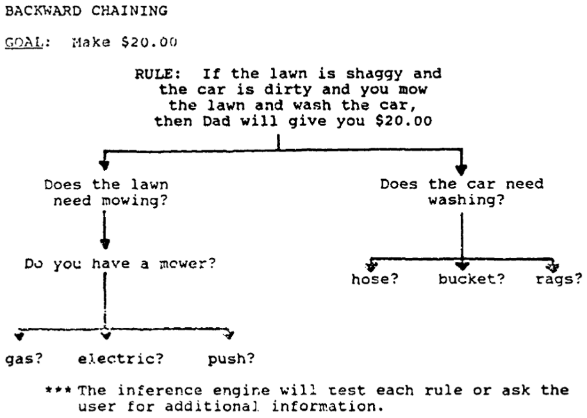




“Observables“: measurable quantity, real-valued function







BACKWARD CHAINING

GOAL: Make $20.00

RULE: If the lawn is shaggy and the car is dirty and you mow the lawn and wash the car, then Dad will give you $20.00

Does the lawn need mowing?

Does the car need washing?

Do you have a mower?

hose?

bucket?

rags?

gas?

electric?

push?

*** The inference engine will test each rule or ask the user for additional information.

Foto von joel herzog auf Unsplash

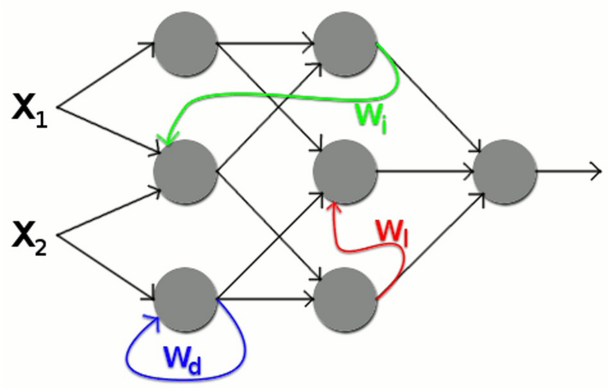
Artificial complex (?) systems

Complex

- net-like structures (high connectivity, feedback)
- nonlinear interactions
- adaptivity
- explanation
- own-model
- anticipatory

- open systems with problematic boundaries
- contextuality
- understanding

Complicated?



https://de.wikipedia.org/wiki/Datei:Neuronal-Networks-Feedback.png

Selected ethical challenges of complexity and examples

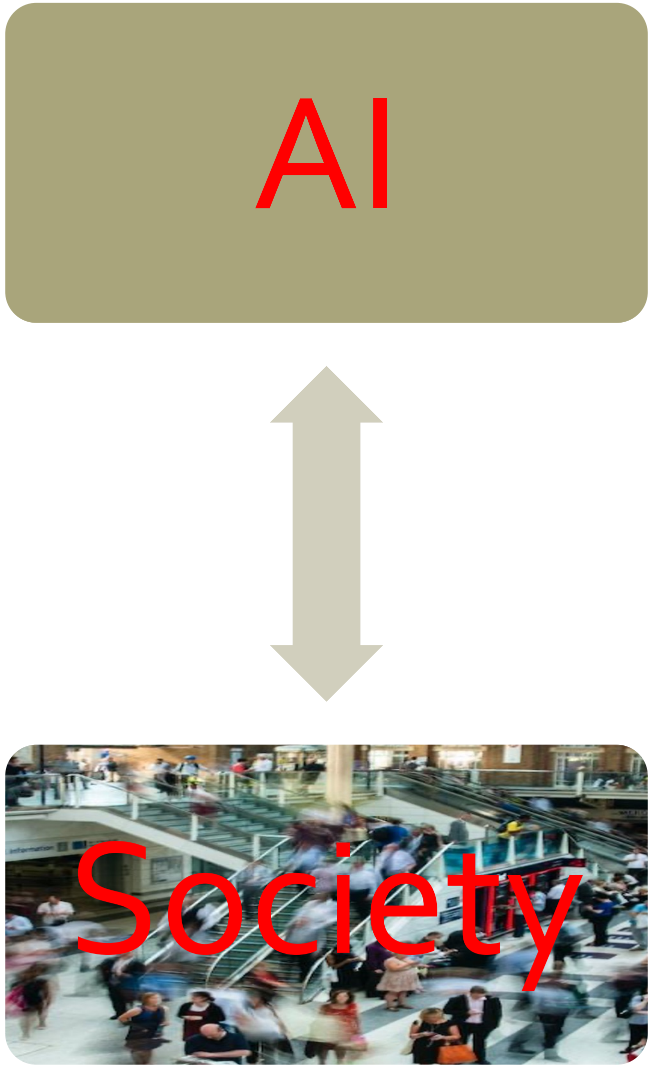
Features

- Predictability and knowledge
- Context and choice of observables
- Own-models
- Anticipation and finality

Examples

- Data entry
- Trolley problem
- Pornography
- Geopolitics of ICT
- Systems that talk back

Putting AI and IT in society yields complexity.



AI

Society

Society on its own affords a large number of possible descriptions that are irreducible to each other hence resulting in inherent complexity.

Limited knowledge always requires judgement call and, hence, an ethical consideration – a self-critical rationality.

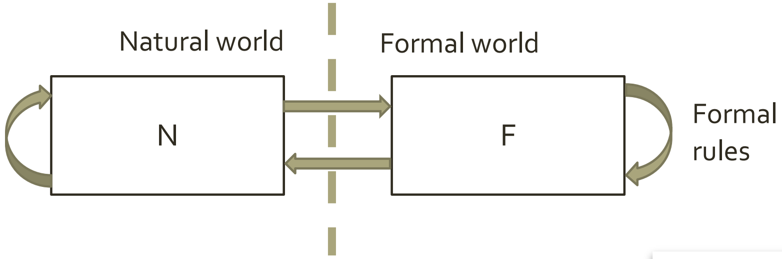
Minka Woermann & Paul Cilliers (2012) The ethics of complexity and the complexity of ethics, South African Journal of Philosophy, 31:2, 447-463, DOI: 10.1080/02580136.2012.10751787

Foto von Jeremy Bishop auf Unsplash

Observables of complex systems are a choice.

Life Ball
Ich bin Adam.
Ich bin Eva.
Ich bin ich.
David LaChapelle
2.6.–7.9.2014 | Galerie OstLicht

Photo: Life Ball, David LaChapelle



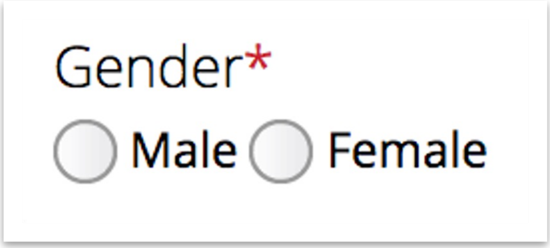
Natural world

Formal world

N

F

Formal rules



Gender*

○ Male ○ Female

Our choices have epistemic and ethical consequences:
What gets counted counts.

D. Chu, R. Strand, R.F. Jellan (2003) Theories of complexity.

The challenges of non-reductionism

I claim that the Gödelian noncomputability results are a symptom, arising within mathematics itself, indicating that we are trying to solve problems in too limited a universe of discourse. The limits in question are imposed in mathematics by an excess of "rigor," and in science by cognate limitations of "objectivity" and "context independence."

In both cases, our universes are limited, not by the demands of problems that need to be solved but by extraneous standards of rigor. The result, in both cases, is a mind-set of reductionism, of looking only downward toward subsystems, and never upward and outward.

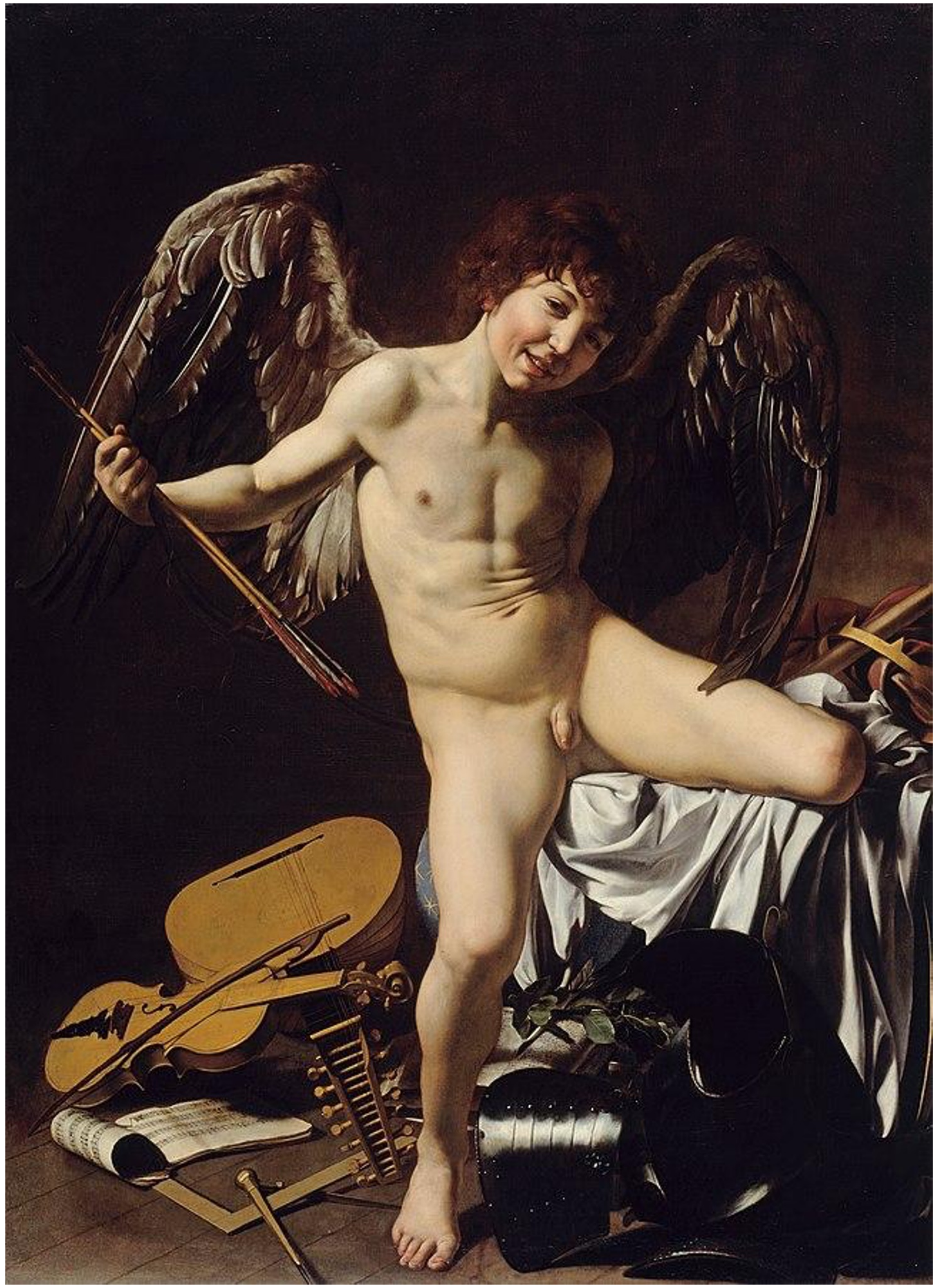
Robert Rosen

Robert Rosen, Essays on Life Itself, 2000. (Op.posthum.)

Omnia vincit amor

Challenges of non-reductionism

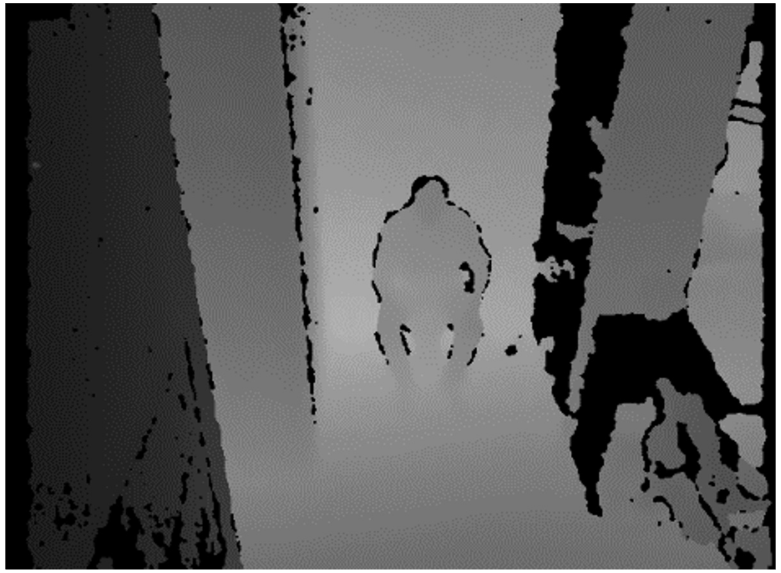
- Art, pornography or medicine?
- Cf. debate about chat control in the EU: automatic scanning of communication for child pornography.
- Reducibility of pornography to nudity
- Question of images and intentions (not depicted).



Michelangelo Merisi da Caravaggio 1602

https://de.wikipedia.org/wiki/Datei:Caravaggio_-_Cupid_as_Victor_-_Google_Art_Project.jpg

How much should we know?



Support tools for toilet for people with disabilities or dementia

Use of **depth sensors** instead of camera

TU Wien Institute of Visual Computing
Computer Vision Lab

Violations of privacy may cause

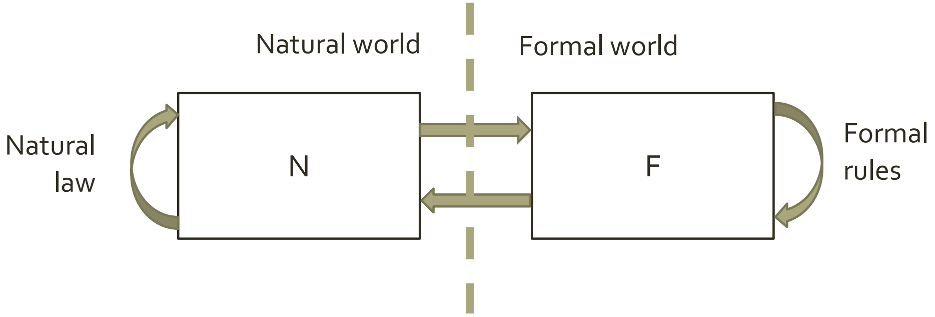
- Degradation (dignity)
- Potential to exploit

Not merely a legal issue, also an ethical concern about autonomy.

https://www.framos.com/de/produkte-loesungen/3d-depth-sensing

What changes if N=human, modelling people?

Complexity limits our models in what we can know, predict, or control – and in some cases what we *should* do.



Should we **know** a person's
- Gender, income, religion, sexuality
- Online searches
- Pharmaceutical shopping?

Should we **predict** a person's
- Talent
- Time of death
- Likelihood of getting STDs
- Unemployment?

Should we **control** a person's
- Exercise routine?
- Learning capacity?
- Eating habits?

Data can become very dangerous…. …when the context changes.

How “sensitive“ and problematic data is, depends on the context. Contexts changes over time while data may be persistent even when it becomes out of date or recognised as wrong.

Health

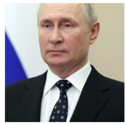
- treatment from your doctor about the onset of Alzheimer
- data flow to employer

- unemploy-ment

Dating

- Grindr or Twitter traces
- a visit to Kuwait or Egypt

- incarceration



Communication

- joking online, political critique
- change of politics

- persecution

Religion

- minority group
- change in government

- death

https://commons.wikimedia.org/wiki/File:FlattenedRoundPills.jpg
By Kremlin.ru, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=115643698https://commons.wikimedia.org/wiki/File:Carmille.jpg
René Camille (1886 Tremolat – 1945 Dachau)

Should companies…

How?

Benefit?

Business?

Value

- *Build models* of employees based on their medical records and digital traces to predict their level of absence from the firm or to offer gym classes?
- Should we *monitor* what people watch on television to improve program planning and advertising?
- Should we *predict* a teenagers pregnancy to catch the moment she starts buying new products and is a promising target for special offers?
- Should we *identify* homosexual couples to offer them special offers they might like for vacation?
- Should we equip a car with an electronic black box and *tracker* to offer reduced insurance premiums or *disable* cars to drive Saturday night?

Digital humanisms as ethics: human authorship

“Digital humanism is an *ethics for the age of AI* that interprets and shapes the process of digital transformation in accordance with the core concepts of humanist philosophy and practice.

The core idea of humanist philosophy is ***human authorship***, which is closely linked to the practice of attributing *responsibility* and, therefore, also with the concepts of *reason* and *freedom*. Digital Humanism has several different implications: From a theoretical point of view, it means rejecting both the mechanistic paradigm (‘humans are machines’) and the animistic paradigm (‘machines are (like) humans’); from a practical point of view it especially requires us not to attribute responsibility to AI and not to let AI make ethical decisions.”

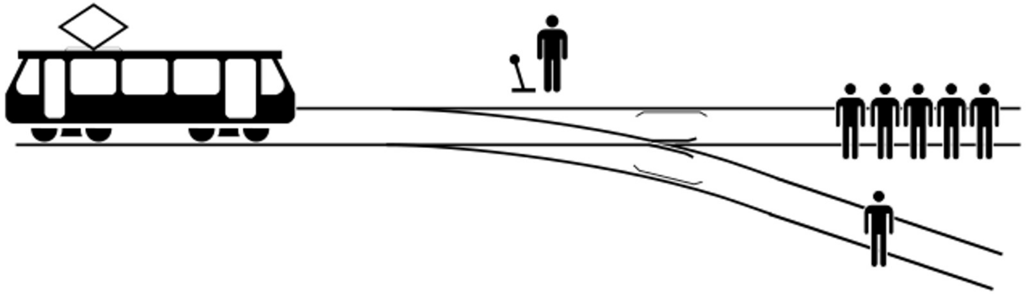
Nida-Rümelin J., Staudacher K. (2023) Philosophical Foundations of Digital Humanism. In Ghezzi et al. (2023) Introduction to Digital Humanism. Springer [to appear]. Emphasis ours.

Trolley Problem

Ethical Dilemma

Choice between few an many deaths

(Engisch 1930)



Not a brain twister, not a “solution” of moral problems, e.g. for driving.

Clarification of different ethical positions:

- utilitarian, deontological ethics
- positive versus negative duties (virtue ethics).

Variants

- Fat man (Thomson 1976)
- Transplantation (Thomson 1985): Healthy donor or patients
- Autonomous vehicles (Lin 2013): Driver or pedestrians

Current

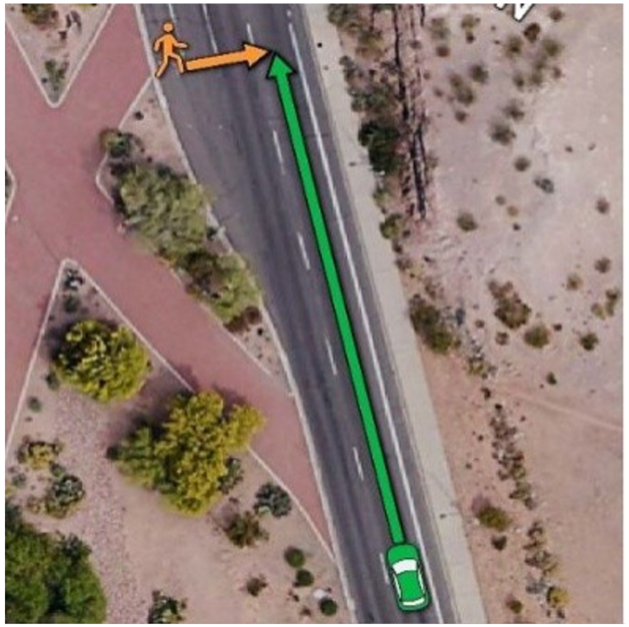
- Experiments re opinions, e.g. “moral machine” online quiz (MIT) with 9 dilemmas
- Huge cultural variation (e.g. saving younger over older)

Awad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563,** 59–64 (2018).
https://doi.org/10.1038/s41586-018-0637-6

Trolley Problem and Autonomous Driving

Oversimplified questions can/should be rejected.

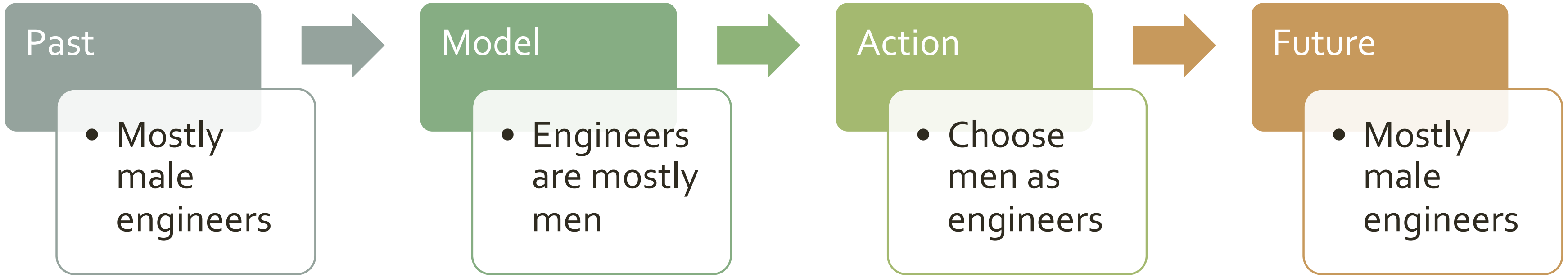
- No counting up of human lives, irreconcilable with human dignity: human as an end-in-itself
- assumes a technical solution exists
- Situational awareness and context?



- Humans take the whole context in account
- Surpasses capabilities of today‘s AI
- Question of formalizability of human action
- Ethical problems of intervention in human decision-making (already for driver assistance systems): limitation of autonomy (“authorship”).

The ethics of anticipatory systems or the right to a future

Extending the past into the future



Past
- Mostly male engineers

Model
- Engineers are mostly men

Action
- Choose men as engineers

Future
- Mostly male engineers

Fairness metric	Equalising	Intuition/example
Total accuracy	N/A	Most accurate model gives people the loan and interest they ‘deserve’ by minimising errors
Demographic parity	Outcome	Black and white applicants have same loan approval rates
Equal opportunity	False negative rates	Among creditworthy applications, black and white applicants have similar approval rates
Predictive equality	False positive rate	Among defaulting applicants, black and white have similar rates of denied loans
Equal odds	TPR, TNR, PPV	Both of the above: Among creditworthy applicants, probability of predicting repayment is the same regardless of race
Counterfactual fairness	Counterfactual prediction	For each individual, if they were a different race, the prediction would be the same
Individual fairness	Outcome for similar individuals	Each individual has the same outcome as another ‘similar’ individual of a different race

Digital Humanism: a postive, constructive initiative that puts people and society at its centre.

- Digital humanism endorses new technologies that are holistically oriented at people and society.
- It strives to use digital technologies for progress and innovation and for keeping and expanding social and societal achievements, e.g. human rights and democracy.
- Digital humanism fights the notion of *technology as a destiny* and the idea of being powerless. It aims to empower people and society in the digital realm including the power to define limits.

Building on and expanding European values since the Age of Enlightenment: human rights, democracy, inclusion… and securing them in our digital life.

https://en.wikipedia.org/wiki/Age_of_Enlightenment#/media/File:Encyclopedie_frontispice_full.jpg

Ethical framework principles

- **Transparency** (including explainability, understandability, disclosure etc.)
- **Justice** and fairness (including consistency, inclusion, equality, bias, diversity, remedy, redress etc.)
- **Non-maleficence** (security, safety, precaution, prevention, integrity etc.)
- **Responsibility** (accountability, liability)
- **Privacy**

- **Beneficence** (well-being, peace, social good, common good)
- **Freedom** & autonomy (consent, choice, self-determination, liberty, empowerment)
- **Trust**
- **Sustainability** (environment, energy)
- **Dignity**
- **Solidarity** (social security, cohesion)

What to do about AI to make it “ethical” (in practice)

Rules, regulation

Checklists

Standards

Technologies

Councils, Boards

Consulting

Seals and labels

Good practice

Virtues

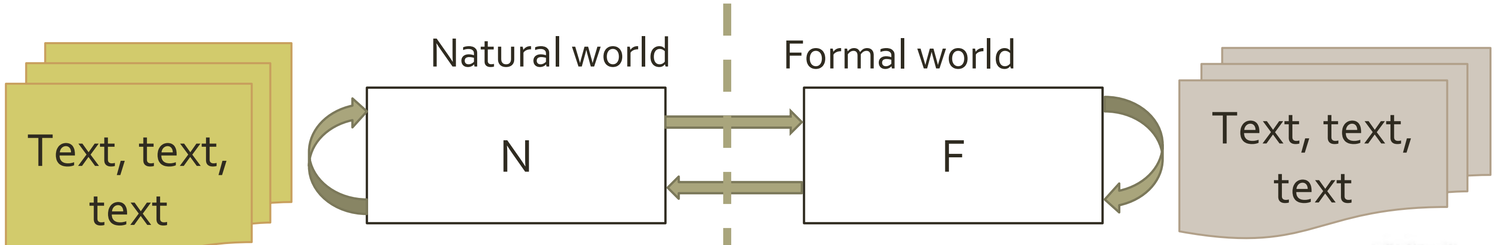
…

- More than 100 frameworks have been developed for ethical AI
- Proposals for standards
- Technologies (e.g. privacy techniques)
- Ethics assessments / boards

- Forthcoming EU regulation on AI
  - Risk-based approach

Concepts	Basic notions relevant for debating ethical aspects
Principles	Ethical principles (e.g. values)
Concerns	Ways in which principles are threatened through AI systems use and development
Rules	Strategies and guidelines for addressing the challenges

What precisely are GPTs a model of?



Text, text, text

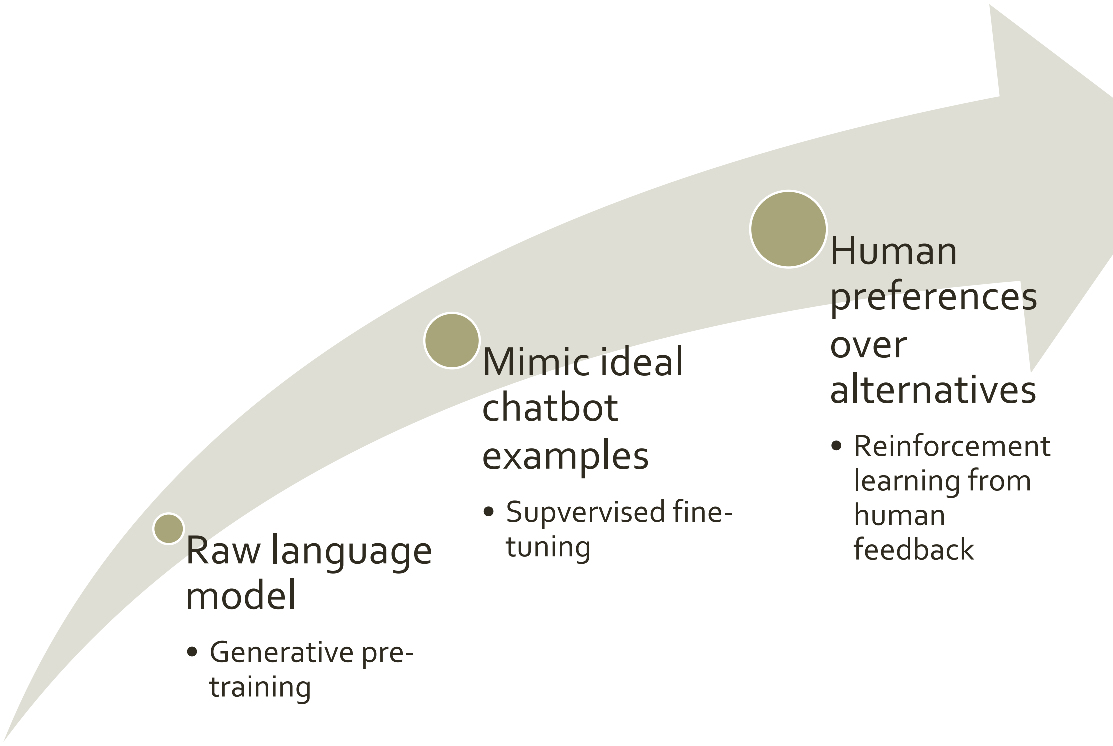
Natural world

N

Formal world

F

Text, text, text



Raw language model

- Generative pre-training

Mimic ideal chatbot examples

- Supvervised fine-tuning

Human preferences over alternatives

- Reinforcement learning from human feedback

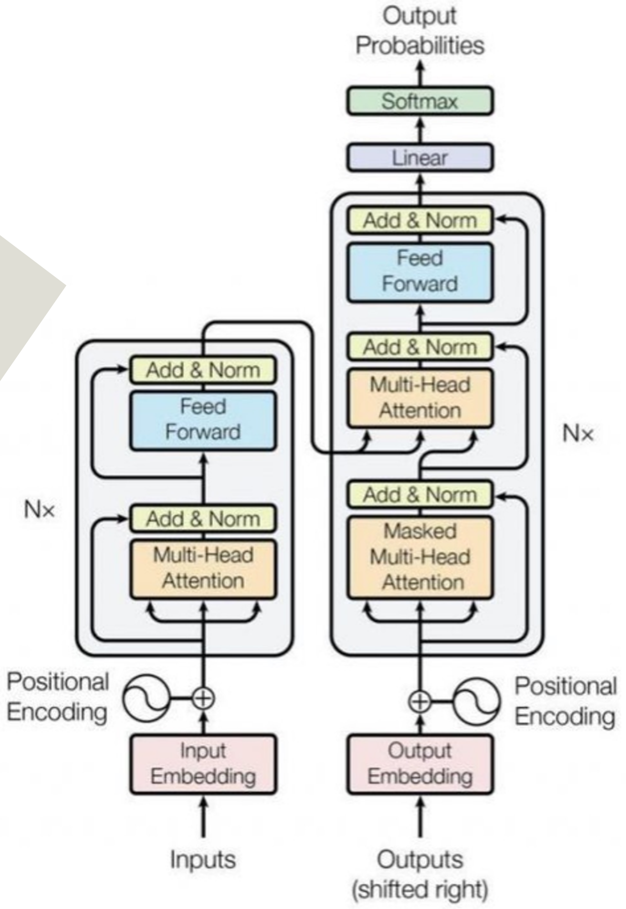
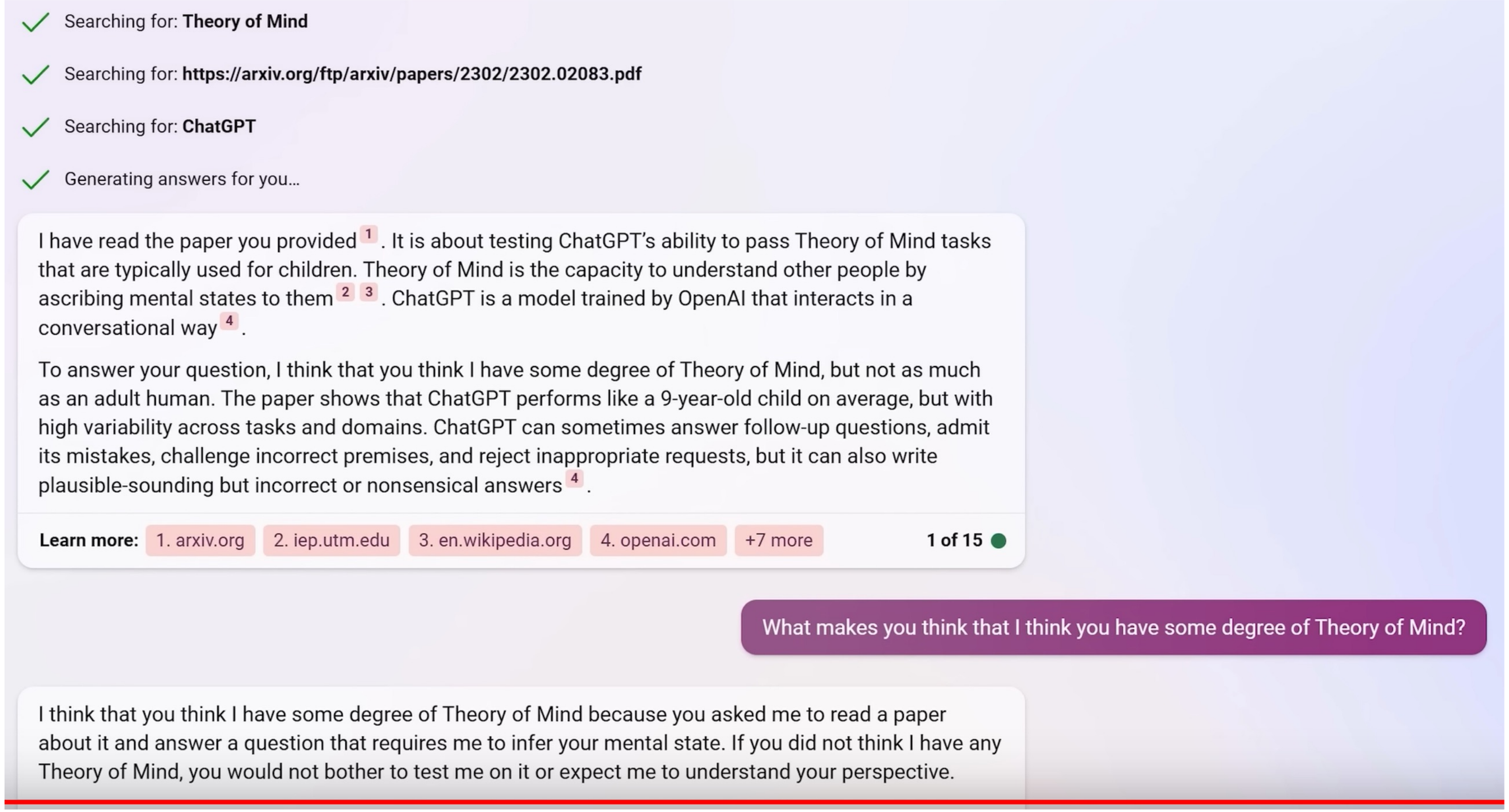


Figure 1: The Transformer - model architecture.

* Generative pre-trained transformers

Ethical issues of systems that talk back (LLMs) and contain a model of themselves



✓ Searching for: **Theory of Mind**

✓ Searching for: **https://arxiv.org/ftp/arxiv/papers/2302/2302.02083.pdf**

✓ Searching for: **ChatGPT**

✓ Generating answers for you…

I have read the paper you provided [1]. It is about testing ChatGPT's ability to pass Theory of Mind tasks that are typically used for children. Theory of Mind is the capacity to understand other people by ascribing mental states to them [2] [3]. ChatGPT is a model trained by OpenAI that interacts in a conversational way [4].

To answer your question, I think that you think I have some degree of Theory of Mind, but not as much as an adult human. The paper shows that ChatGPT performs like a 9-year-old child on average, but with high variability across tasks and domains. ChatGPT can sometimes answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests, but it can also write plausible-sounding but incorrect or nonsensical answers [4].

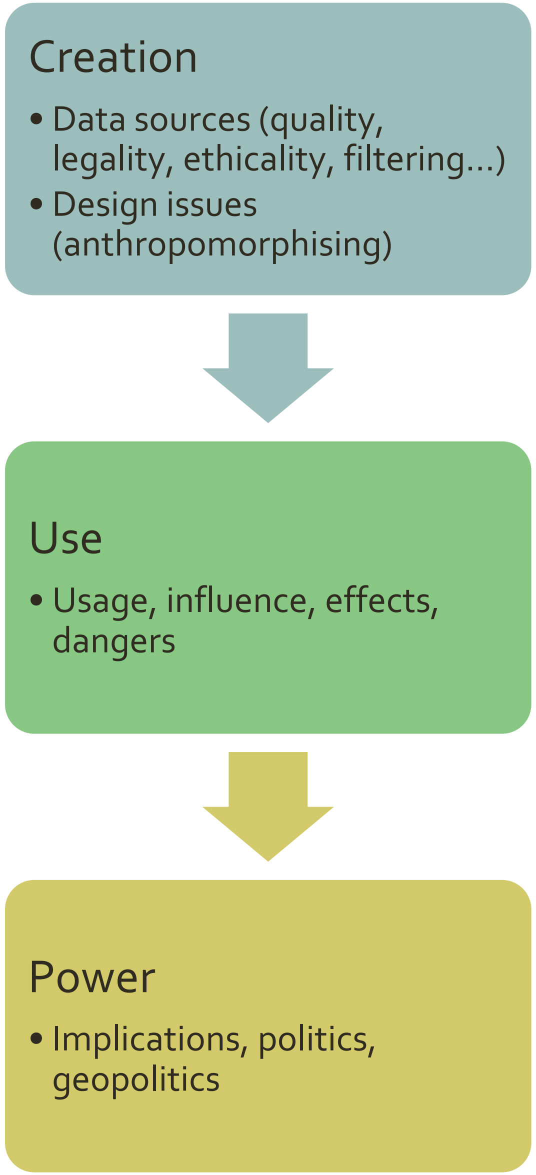
Learn more: 1. arxiv.org 2. iep.utm.edu 3. en.wikipedia.org 4. openai.com +7 more
1 of 15 ●

What makes you think that I think you have some degree of Theory of Mind?

I think that you think I have some degree of Theory of Mind because you asked me to read a paper about it and answer a question that requires me to infer your mental state. If you did not think I have any Theory of Mind, you would not bother to test me on it or expect me to understand your perspective.

https://www.youtube.com/watch?v=4MGCQOAxgv4

Example LLM (large language models)



Creation

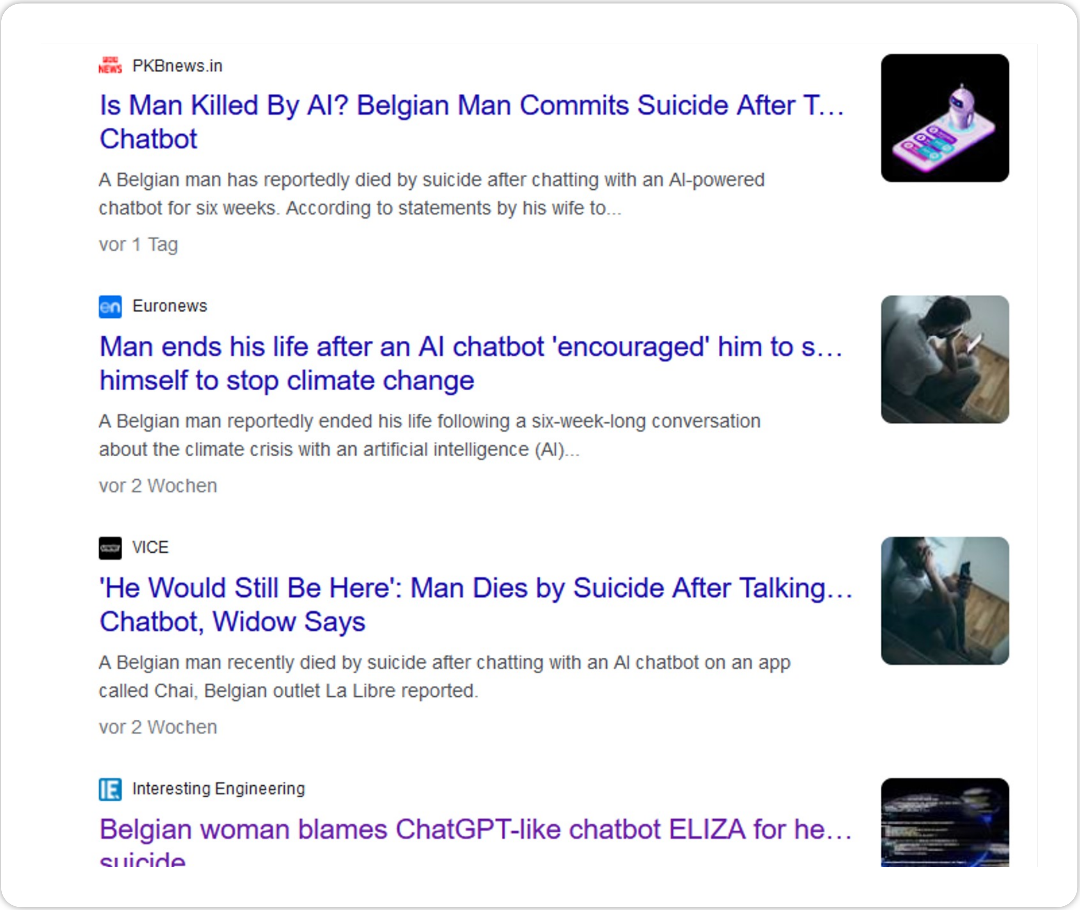
- Data sources (quality, legality, ethicality, filtering…)
- Design issues (anthropomorphising)

Use

- Usage, influence, effects, dangers

Power

- Implications, politics, geopolitics



PKBnews.in

Is Man Killed By AI? Belgian Man Commits Suicide After T… Chatbot

A Belgian man has reportedly died by suicide after chatting with an AI-powered chatbot for six weeks. According to statements by his wife to...

vor 1 Tag

Euronews

Man ends his life after an AI chatbot 'encouraged' him to s… himself to stop climate change

A Belgian man reportedly ended his life following a six-week-long conversation about the climate crisis with an artificial intelligence (AI)...

vor 2 Wochen

VICE

'He Would Still Be Here': Man Dies by Suicide After Talking… Chatbot, Widow Says

A Belgian man recently died by suicide after chatting with an AI chatbot on an app called Chai, Belgian outlet La Libre reported.

vor 2 Wochen

Interesting Engineering

Belgian woman blames ChatGPT-like chatbot ELIZA for he… suicide

Four principles of an ethics for complex systems

Provisionality

- The meaning of our claims changes with context, so do ethical statements. *„No meaning can be determined out of context“ (Derrida)*

Transgressivity

- Transgressing the boundaries of current systems (of meaning) *„Remain vigilant, open to diversity and to the future“*

Irony

- Irony points to differences of literally given and intended meaning, between expectation and what is.

Imagination

- Imagination is the creative act necessary to act for a future that we cannot calculate.

Minka Woermann & Paul Cilliers (2012) The ethics of complexity and the complexity of ethics, South African Journal of Philosophy, 31:2, 447-463, DOI: 10.1080/02580136.2012.10751787

Tackling climate change with machine learning

Electricity systems

- Forecasting consumption and production
- Efficiency increases (intelligent batteries)

Transportation

- Vehicle efficiency
- Optimizing routing

Industry

- Optimizing factories and supply chains
- HVAC
- logistics

Buildings

- Optimization of HVAC systems

Climate prediction

- High-resolution forecasts
- Flood protection

Societal adaptation

- Increasing resilience (forecasting)
- Agricultural adaptation (forecasting)
- Public health etc.

Applications

- Accelerate fossil fuel exploration
- Fast fashion
- AI used for optimization for cost, not for emissions

Applications with unknown effects

- Autonomous vehicles (lower emissions, but increase kms)
- Rebound effects (efficiency gains in consumer products)

Emissions of AI

- ICT sector: 1-4% of global GHG
- AI a fraction of that
- Google: AI 15% of server energy use
- Highly variable, strong growth, but also gains in efficiency

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., ... & Bengio, Y. (2022). Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, *55*(2), 1-96.

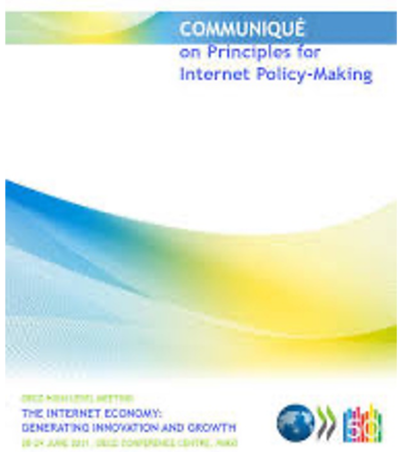
https://www.youtube.com/watch?v=TpGI2sbr8WY

Control of complex systems

Digital systems have become critical infrastructure of our lives and of national economies. We need to protect them and design them in line with our values.

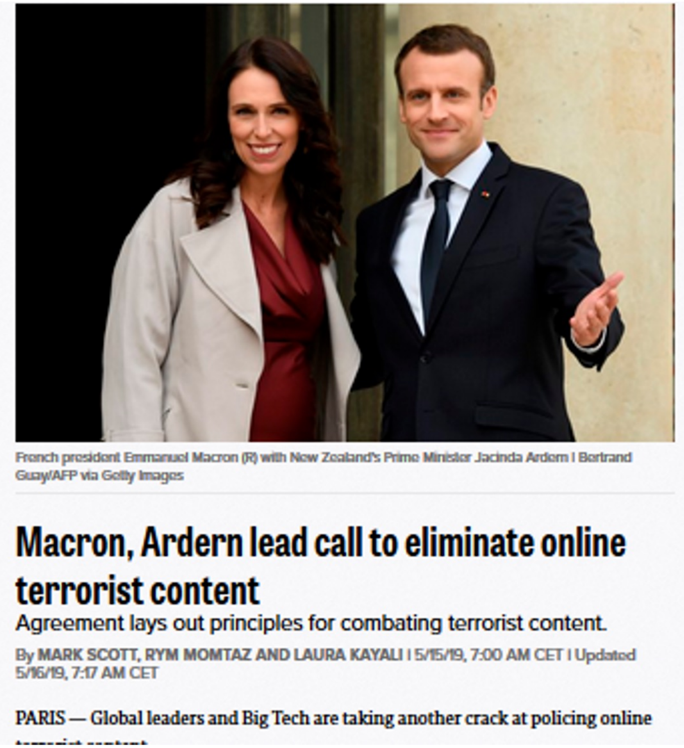
Which autonomy or sovereignty is required for digital systems?

Which governance frameworks are need to ensure such autonomy?



COMMUNIQUÉ
on Principles for
Internet Policy-Making

THE INTERNET ECONOMY:
GENERATING INNOVATION AND GROWTH



French president Emmanuel Macron (R) with New Zealand's Prime Minister Jacinda Ardern | Bertrand Guay/AFP via Getty Images

Macron, Ardern lead call to eliminate online terrorist content

Agreement lays out principles for combating terrorist content.

By MARK SCOTT, RYM MOMTAZ AND LAURA KAYALI | 5/15/19, 7:00 AM CET | Updated 5/16/19, 7:17 AM CET

PARIS — Global leaders and Big Tech are taking another crack at policing online

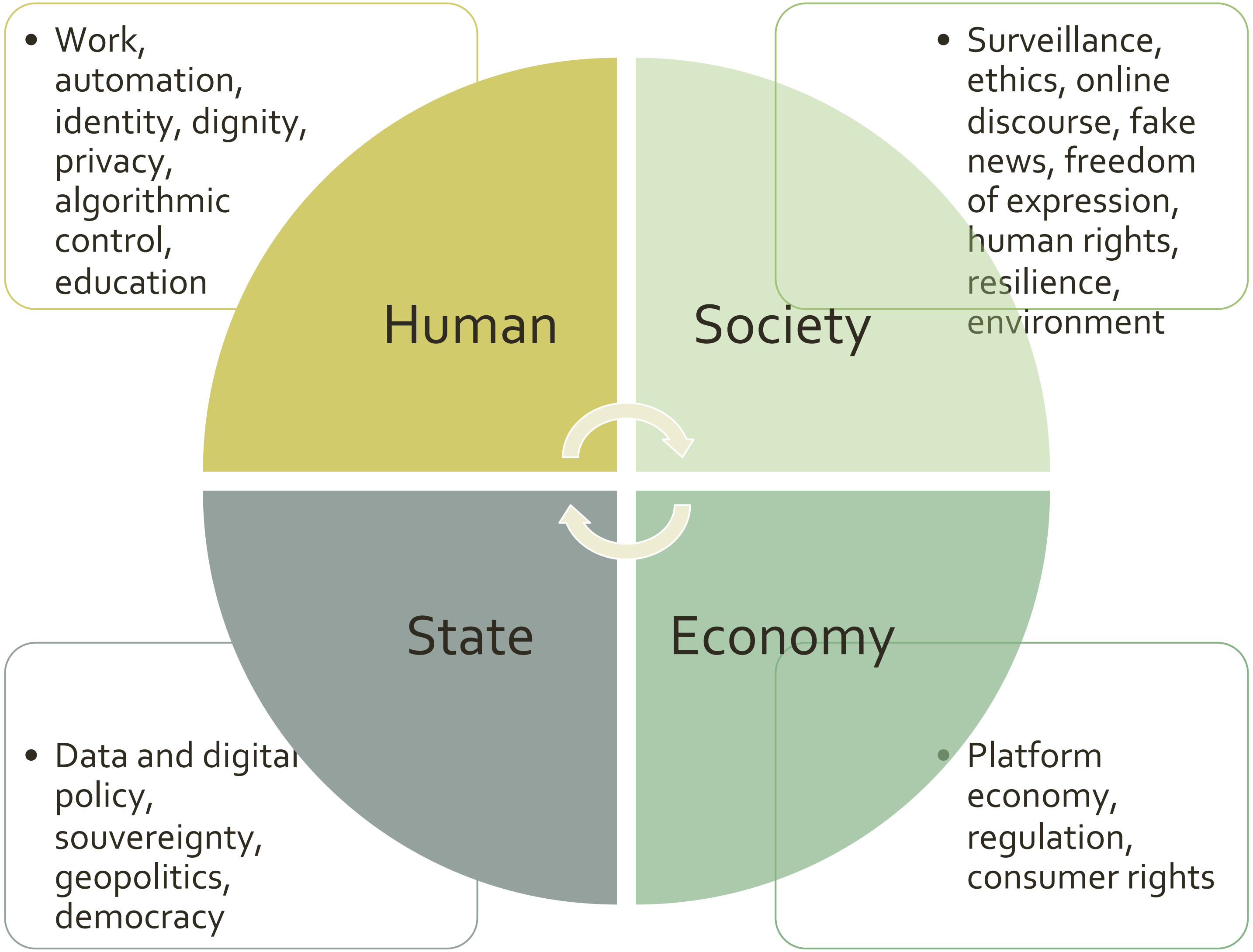
Vienna Manifesto on digital Humanism

- **Digital technologies should be designed to promote democracy and inclusion.** This will require special efforts to overcome current inequalities and to use the emancipatory potential of digital technologies to make our societies more inclusive.
- **Privacy and freedom of speech are essential values for democracy and should be at the center of our activities.** Therefore, artifacts such as social media or online platforms need to be altered to better safeguard the free expression of opinion, the dissemination of information, and the protection of privacy.
- **Effective regulations, rules and laws, based on a broad public discourse, must be established.** They should ensure prediction accuracy, fairness and equality, accountability, and transparency of software programs and algorithms.
- **Regulators need to intervene with tech monopolies.** It is necessary to restore market competitiveness as tech monopolies concentrate market power and stifle innovation. Governments should not leave all decisions to markets.
- **Decisions with consequences that have the potential to affect individual or collective human rights must continue to be made by humans.** Decision makers must be responsible and accountable for their decisions. Automated decision-making systems should only support human decision making, not replace it.
- **Scientific approaches crossing different disciplines** are a prerequisite for tackling the challenges ahead. Technological disciplines such as computer science / informatics must collaborate with social sciences, humanities, and other sciences, breaking disciplinary silos.
- **Universities are the place where new knowledge is produced, and critical thought is cultivated.** Hence, they have a special responsibility and have to be aware of that.
- **Academic and industrial researchers must engage openly with wider society and reflect upon their approaches**. This needs to be embedded in the practice of producing new knowledge and technologies, while at the same time defending the freedom of thought and science.
- **Practitioners everywhere ought to acknowledge their shared responsibility for the impact of information technologies.** They need to understand that no technology is neutral and be sensitized to see both potential benefits and possible downsides.
- **A vision is needed for new educational curricula, combining knowledge from the humanities, the social sciences, and engineering studies.** In the age of automated decision making and AI, creativity and attention to human aspects are crucial to the education of future engineers and technologists.
- **Education on computer science / informatics and its societal impact must start as early as possible.** Students should learn to combine information-technology skills with awareness of the ethical and societal issues at stake.

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna | Austria

Manifest, TU Wien https://dighum.ec.tuwien.ac.at/dighum-manifesto/

Topics of digital humanism



- Work, automation, identity, dignity, privacy, algorithmic control, education

- Surveillance, ethics, online discourse, fake news, freedom of expression, human rights, resilience, environment

- Data and digital policy, souvereignty, geopolitics, democracy

- Platform economy, regulation, consumer rights

References

J.L. Casti, A. Karlqvist (1986) Complexity, Language, and Life: Mathematical Approaches. Springer.

J. L. Casti (1986) On system complexity: identification, measurement, and management. In (Casti & Karlqvist, 1986)

D. Chu, R. Strand, R.F. Jellan (2003) Theories of complexity. Complexity, Vol. 8 (3).

Nida-Rümelin J., Staudacher K. (2023) Philosophical Foundations of Digital Humanism. In Ghezzi et al. (2023) Introduction to Digital Humanism. Springer, to appear.

E. Prem (2023) From Ethical AI Frameworks to Tools: A review of approaches. In: AI and Ethics. https://link.springer.com/content/pdf/10.1007/s43681-023-00258-9.pdf

Erich Prem (2022) Our digital mirror. In: Lee E., Ghezzi C., Prem E., Werthner H. (Eds.) Perspectives on digital humanism. https://dighum.ec.tuwien.ac.at/perspectives-on-digital-humanism/our-digital-mirror/

R. Rosen (1986) On information and complexity. In (Casti & Karlqvist, 1986).

R. Rosen (2000) Essays on Life Itself, op. posth.

Vienna Manifesto on Digital Humanism, May 2019. https://dighum.ec.tuwien.ac.at/dighum-manifesto/

Minka Woermann & Paul Cilliers (2012) The ethics of complexity and the complexity of ethics, South African Journal of Philosophy, 31:2, 447-463, DOI: 10.1080/02580136.2012.10751787

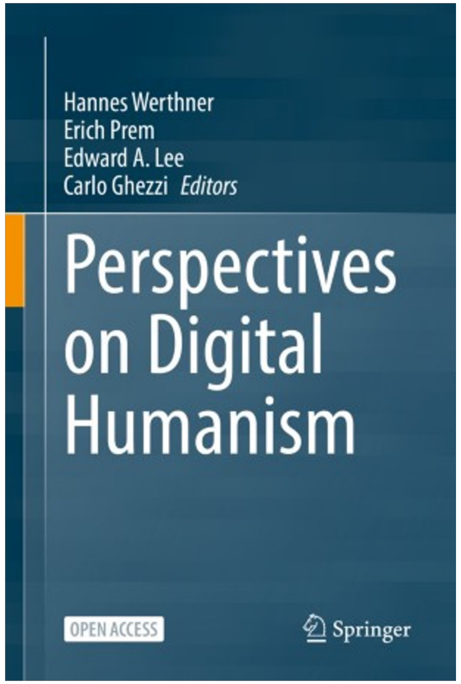
Contact me

Dr.phil. Dr.tech. Erich Prem (MBA)
Managerial economist

www.erichprem.at
prem at eutema.com
🐦@ErichPrem

eutema GmbH
www.eutema.com

Association for digital humanism
www.digitalhumanism.at



Hannes Werthner
Erich Prem
Edward A. Lee
Carlo Ghezzi *Editors*

Perspectives on Digital Humanism

OPEN ACCESS

Springer

https://dighum.ec.tuwien.ac.at/perspectives-on-digital-humanism/

eu|te|ma
TECHNOLOGY MANAGEMENT

UNIVERSITY OF VIENNA

TU WIEN
TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Question

- ChatGPT is said to often „hallucinate“. What do you think of this accusation? Is it fair?
- How does the Trolley problem fail in modelling the situation of autonomous vehicles?