MDU 2024 09 30

AI TECHNOLOGY ETHICAL BY DESIGN?

Intelligent technology, promises, and challenges

Gordana Dodig Crnkovic

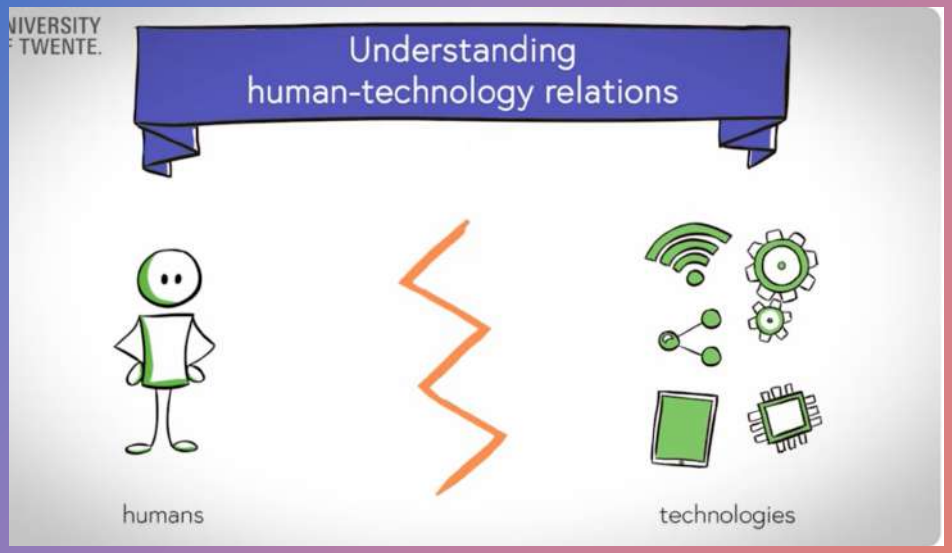
Professor of Computer Science

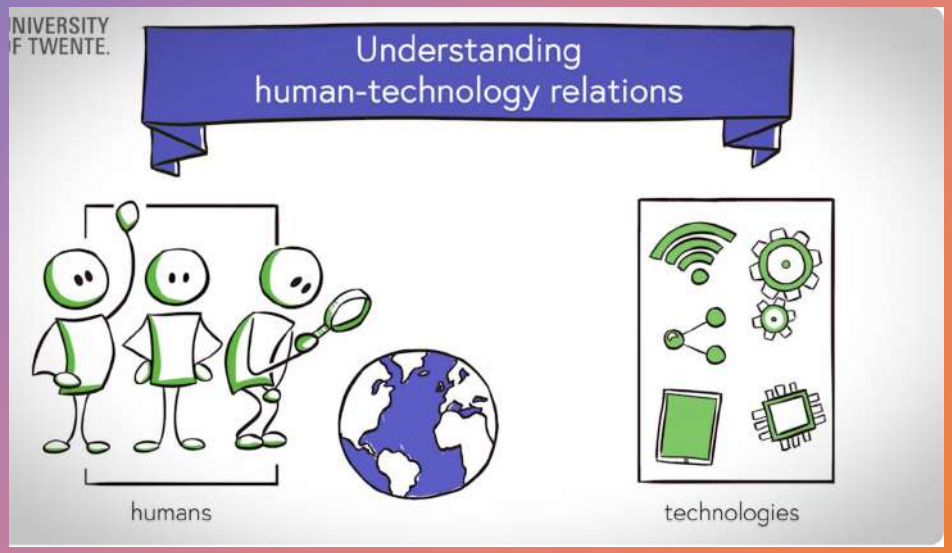
Mälardalen University, Sweden &

Chalmers Technical University | University of Gothenburg

http://gordana.se/

http://www.gordana.se/work/presentations.html





Design = "Doing ethics by other means" Peter-Paul Verbeek

In a similar way as observations are "theory-laden" (Thomas Kuhn), the designed artifacts (technological products) are "value-laden".

https://www.youtube.com/watch?v=FVhrLwBNbvU Peter-Paul Verbeek Explaining Technological Mediation

Gordana Dodig-Crnkovic, affiliations

GÖTEBORGS
UNIVERSITET

School of Innovation, Design and Engineering

Division of Computer Science and Software Engineering

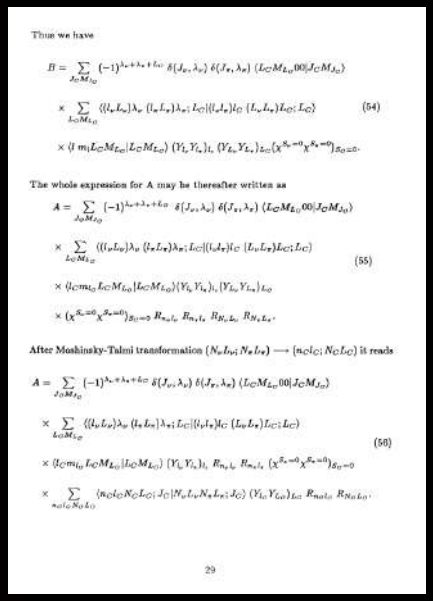
Research groups:
Artificial Intelligence and Intelligent Systems
Ubiquitous Computing

Department of Computer Science and Engineering

Division:
Computer Science and Software Engineering

Research groups:
Interaction Design and Software Engineering
Critical Robotics

My background - from formal to natural languages



Thus we have

$$B = \sum_{J_C M_{J_C}} (-1)^{\lambda_\sigma+\lambda_\pi+L_C} \delta(J_\sigma,\lambda_\sigma)\,\delta(J_\pi,\lambda_\pi)\,\langle L_C M_{L_C} 00|J_C M_{J_C}\rangle$$

$$\times \sum_{L_C M_{L_C}} \langle\{[l_\sigma L_\sigma]\lambda_\sigma\,[l_\pi L_\pi]\lambda_\pi\}_{L_C}|[[l_\sigma l_\pi]_{l_C}\,[L_\sigma L_\pi]_{L_C}]_{J_C}\rangle \qquad (54)$$

$$\times \langle l\ m_l L_C M_{L_C}|L_C M_{L_C}\rangle\,\{Y_{l_\sigma} Y_{L_\sigma}\}_{l_\sigma}\,\{Y_{l_\pi} Y_{L_\pi}\}_{l_\pi}\,\{\chi^{S_\sigma=0}\chi^{S_\pi=0}\}_{S_C=0}.$$

The whole expression for A may be thereafter written as

$$A = \sum_{J_C M_{J_C}} (-1)^{\lambda_\sigma+\lambda_\pi+L_C}\ \delta(J_\sigma,\lambda_\sigma)\,\delta(J_\pi,\lambda_\pi)\,\langle L_C M_{L_C} 00|J_C M_{J_C}\rangle$$

$$\times \sum_{L_C M_{L_C}} \langle([l_\sigma L_\sigma]\lambda_\sigma\,[l_\pi L_\pi]\lambda_\pi;L_C|[[l_\sigma l_\pi]_{l_C}\,[L_\sigma L_\pi]_{L_C};L_C\rangle \qquad (55)$$

$$\times \langle l_C m_{l_C} L_C M_{L_C}|L_C M_{L_C}\rangle\{Y_{l_\sigma} Y_{l_\pi}\}_{l_C}\{Y_{L_\sigma} Y_{L_\pi}\}_{L_C}$$

$$\times (\chi^{S_\sigma=0}\chi^{S_\pi=0})_{S_C=0}\,R_{n_\sigma l_\sigma}\,R_{n_\pi l_\pi}\,R_{N_\sigma L_\sigma}\,R_{N_\pi L_\pi}.$$

After Moshinsky-Talmi transformation $\{N_\sigma l_\sigma; N_\pi L_\pi\} \longrightarrow \{n_C l_C; N_C L_C\}$ it reads

$$A = \sum_{J_C M_{J_C}} (-1)^{\lambda_\sigma+\lambda_\pi+L_C}\,\delta(J_\sigma,\lambda_\sigma)\,\delta(J_\pi,\lambda_\pi)\,\langle L_C M_{L_C} 00|J_C M_{J_C}\rangle$$

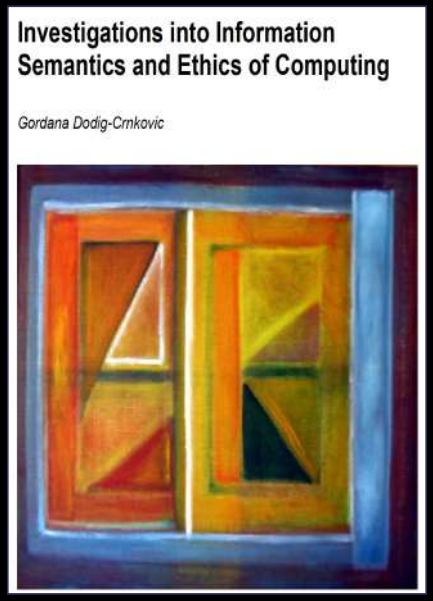
$$\times \sum_{L_C M_{L_C}} \langle([l_\sigma L_\sigma]\lambda_\sigma\,[l_\pi L_\pi]\lambda_\pi;L_C|[[l_\sigma l_\pi]_{l_C}\,[L_\sigma L_\pi]_{L_C};L_C\rangle \qquad (56)$$

$$\times \langle l_C m_{l_C} L_C M_{L_C}|L_C M_{L_C}\rangle\,\{Y_{l_\sigma} Y_{l_\pi}\}_{l_C}\,R_{n_\sigma l_\sigma}\,R_{n_\pi l_\pi}\,\{\chi^{S_\sigma=0}\chi^{S_\pi=0}\}_{S_C=0}$$

$$\times \sum_{n_C l_C N_C L_C} \langle n_C l_C N_C L_C; J_C|N_\sigma L_\sigma N_\pi L_\pi; J_C\rangle\,\{Y_{l_C} Y_{L_C}\}_{L_C}\,R_{n_C l_C}\,R_{N_C L_C}.$$

29

PhD in Physics, 1988
On Alpha-decay, Department of Physics, University of Zagreb



PhD in Computing, 2006
Computer Science, Mälardalen University

Current: Morphological Cognitive and Intelligent Computing, AI Ethics, Digital Ethics, Digital Humanism

https://www.gordana.se/work/courses.html

https://tinyurl.com/34r7xxyw

The Perspective

The aim of this lecture – to offer new views

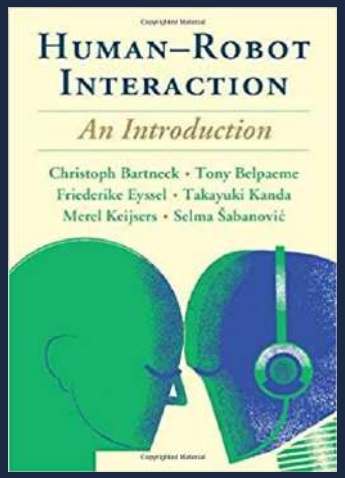
As the topic of Design Ethics, AI ethics and even AV ethics are huge, what this lecture can do is to open the window with a view, giving you just a glimpse of a huge unexplored territory in front of us.

”I invite readers not on a visit to an archaeological museum, but rather on an adventure in science in making”

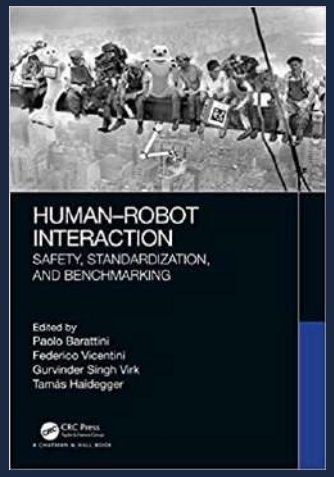
Ilya Prigogine. The End of Certainty: Time, Chaos and New Laws of Nature, 1997

https://www.onventanas.com/historia-vidrio/ventana-japonesa/#iLightbox[postimages]/0

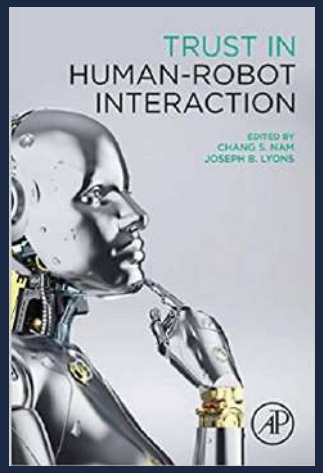
Value-Centric Design for Robot-Human Interactions



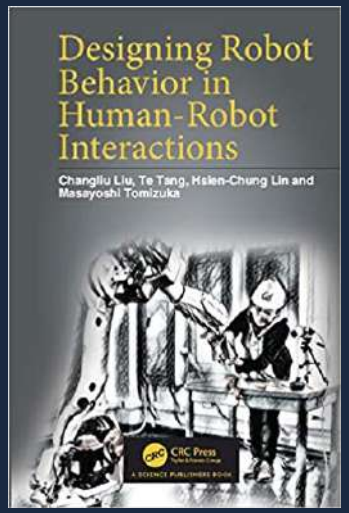
https://www.amazon.com/Human-Robot-Interaction-Introduction-Christoph-Bartneck/dp/1108735401?asin=B0845ZM9M4&revisionId=fixed_format&format=4&depth=2



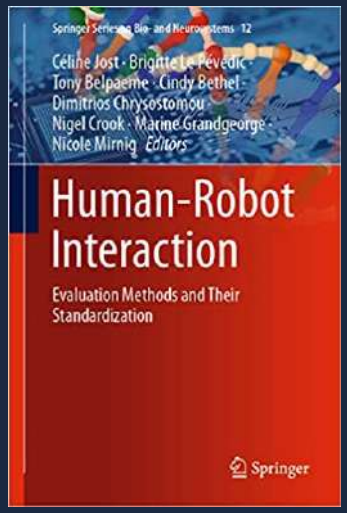
https://www.amazon.com/Human-Robot-Interaction-Safety-Standardization-Benchmarking-ebook-dp-B07QM9VTR8/dp/B07QM9VTR8/ref=mt_other?_encoding=UTF8&me=&qid=1628325169



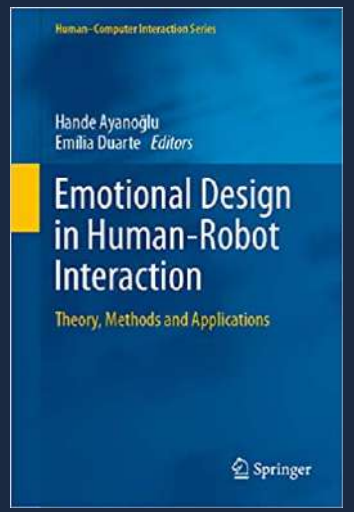
https://www.amazon.com/Trust-Human-Robot-Interaction-Chang-Nam-ebook/dp/B08NVM7SGF/ref=sr_1_3?dchild=1&keywords=Human-Robot+Interaction&qid=1628325169&s=books&sr=1-3



https://www.amazon.com/Designing-Robot-Behavior-Human-Robot-Interactions/dp/0367179695/ref=sr_1_4?dchild=1&keywords=Human-Robot+Interaction&qid=1628325169&s=books&sr=1-4



https://www.amazon.com/Human-Robot-Interaction-Evaluation-Standardization-Neurosystems-ebook/dp/B088LYNPYB/ref=sr_1_7?dchild=1&keywords=Human-Robot+Interaction&qid=1628325169&s=books&sr=1-7



https://www.amazon.com/Emotional-Design-Human-Robot-Interaction-Human-Computer-ebook-dp-B07FR6QZPQ/dp/B07FR6QZPQ/ref=mt_other?_encoding=UTF8&me=&qid=1628325169

Developing intelligent robots that we can trust and like

Developing intelligent autonomous robots that we can trust and enjoy presupposes they meet our expectation on values with anticipated beneficial influence on the societies and individuals, globally with respect to **ELSI (Ethical, Legal and Social Implications)**.

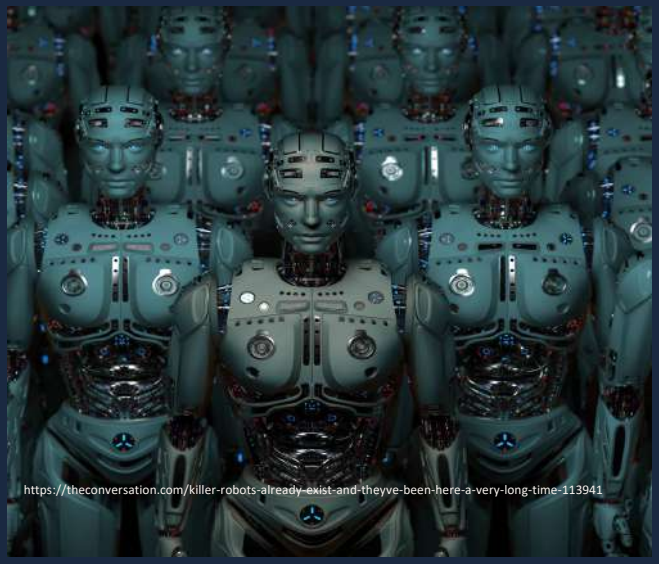
Values with questions of good and bad, right and wrong, and values, in general, are studied within the field of ethics.

The emerging fields of Robotic Ethics, AI ethics and specifically ethics of intelligent autonomous robotic cars are good examples of ethics research with actionable practical value.

In those ethical fields, a variety of stakeholders, including the legal system with other societal and governmental actors, companies and businesses, collaborate bringing about shared view of ELSI.

Drawing from the existing literature on ethics of AI and robotics, and our work on autonomous intelligent robocars, our contribution consists in lessons learned for ethics of autonomous intelligent robots in general.

https://robohub.org/from-disembodied-bytes-to-robots-that-think-and-act-like-humans/



https://theconversation.com/killer-robots-already-exist-and-theyve-been-here-a-very-long-time-113941

Framtiden är inte oundviklig, utan vår att skapa!

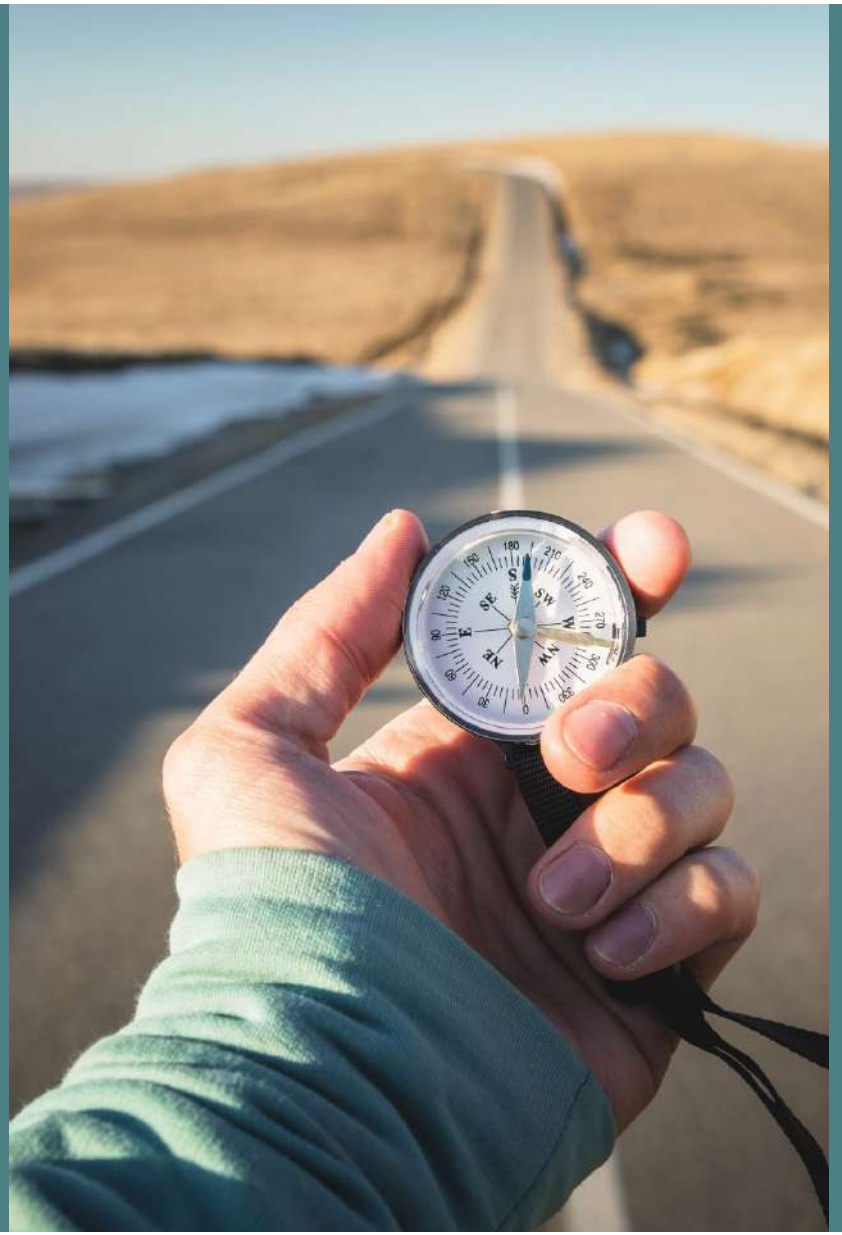
The future is not inevitable, but ours to create!

"AI is not a done deal. We're building the road, as we walk it, and we can collectively decide what direction we want to go in, together."

"AI är inte en färdig affär. Vi bygger vägen, medan vi går den, och vi kan tillsammans bestämma vilken riktning vi vill gå i, tillsammans."

- Sasha Luccioni, 2023

https://www.nationalacademies.org/news/2024/02/just-how-intelligent-is-artificial-intelligence



Ethics as a Participatory and Iterative Process

Etik som en deltagarstyrd och iterativ process

Ethics involves a participatory and iterative process of ethical reflection, inquiry, and deliberation. Combining action and reflection is crucial.

Etik innebär en deltagande och iterativ process av etisk reflektion, undersökning och överläggning. Att kombinera handling och reflektion är avgörande.

It is instructive to go back and forth between zooming out and zooming in on the problem.

Det är lärorikt att gå fram och tillbaka mellan att zooma ut och zooma in på problemet.

In this process, we consult different ethical approaches (Consequentialism, Duty ethics, Virtue ethics, Relational ethics, etc.)

I denna process använder vi olika etiska förhållningssätt (konsekventialism, pliktetik, dygdetik, relationsetik, etc.)

Dygdetik (Virtue Ethics) är en moralfilosofisk teori som fokuserar på individens karaktär och dygder som grund för att bedöma etiskt handlande

Pliktetik (Duty Ethics) betonar regler och plikter

Konsekventialism (Consequentialism) bedömer handlingar utifrån deras konsekvenser.

Relationell etik (Relational Ethics) fokuserar på betydelsen av relationer och kontext i moraliskt handlande.

Methods from Human-Centered Design (HCD) organizing participatory and iterative processes, Value Sensitive Design (VSD), bringing different stakeholder values, and Responsible Innovation (RI) with a focus on inclusion, participation, and diversity.

Metoder från Human-Centered Design (HCD) som organiserar deltagande och iterativa processer, Value Sensitive Design (VSD), som ger olika intressentvärden och Responsible Innovation (RI) med fokus på inkludering, delaktighet och mångfald.

We face complex, interdisciplinary, and global challenges: climate crisis, political polarization, and inequalities. These are all **wicked problems**, which require diverse disciplines, both to better understand the problem and to envision and create solutions.

Vi står inför komplexa, tvärvetenskapliga och globala utmaningar: klimatkris, politisk polarisering och ojämlikheter. Dessa är alla onda problem, som kräver olika discipliner, både för att bättre förstå problemet och för att föreställa sig och skapa lösningar.

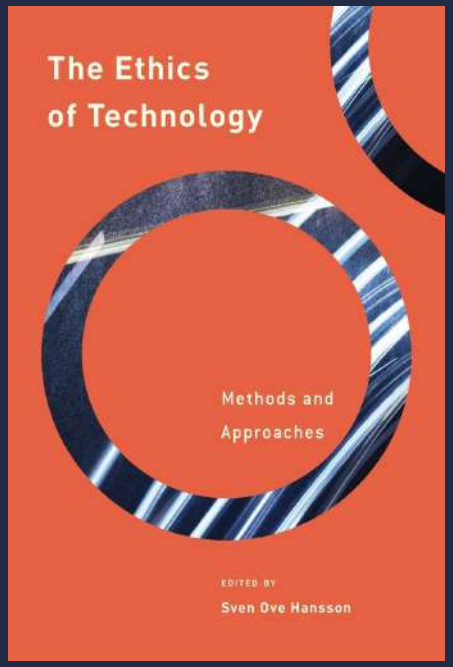
Doing ethics is not always easy or pleasant. It can involve asking uneasy questions, creating awkward situations, and tolerating tension and uncertainty.

Att använda etik är inte alltid lätt eller trevligt. Det kan innebära att ställa obekväma frågor, skapa besvärliga situationer och tolerera spänningar och osäkerhet.

Stakeholders = aktörer, intressenter (aktiva och passiva)

https://dl.acm.org/doi/pdf/10.1145/3550069 Marc Steen

Innehåll



Sven Ove Hansson, professor i filosofi vid Kungliga Tekniska högskolan (KTH). Medicine kandidat vid Lunds universitet, filosofiekandidat vid Uppsala universitet, och disputerade i teoretisk filosofi, också vid Uppsala universitet. Han avlade en andra doktorsexamen, i praktisk filosofi vid Lunds universitet.

Några av AI:s etiska utmaningar

"Ethical guidelines on the use of AI and data in teaching and learning are an incremental process of continuous deliberation and learning."

Expert Group on AI and data in education and training

Med tanke på att det krävs en stor mängd data för att träna AI-systemen, algoritmerna till sin natur är automatiserade och deras applikationer går att skala upp innebär användningen av AI att viktiga frågor måste diskuteras när det gäller personuppgifter, dataskydd och integritet.

Dessa etiska riktlinjer för användningen av AI och data vid undervisning och inlärning har utformats på ett sätt som ska hjälpa lärare att förstå vilken potential användningen av AI-applikationer och data kan ha inom utbildningen och öka medvetenheten om möjliga risker. De kan då inta ett positivt, kritiskt och etiskt förhållningssätt till AI-systemen och förverkliga deras fulla potential.

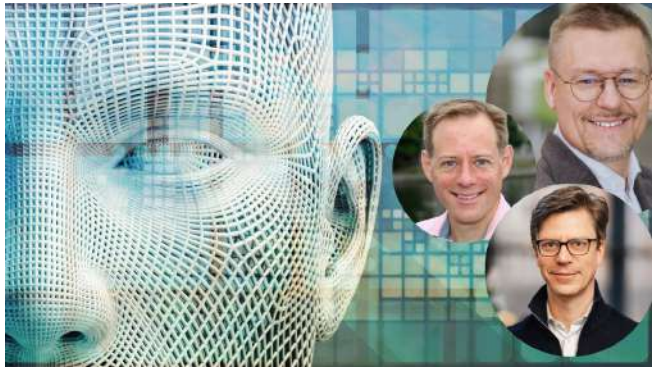
Transparens och förklarbarhet i AI-system

Bias (fördomar) och rättvisa i AI-algoritmer

Integritet och dataskydd

Ansvarsfrågor vid AI-baserade beslut

AI och samhället



https://www.dagenssamhalle.se/samhalle-och-valfard/digitalisering/experterna-se-upp-for-farorna-med-ai-stora-olyckor/ Fredrik Heintz, Daniel Gillblad och Magnus Mähring, Experterna: Se upp för farorna med AI

- Farorna med nya tekniken.
- Så kan riskerna förebyggas.
- Då kan AI-modellerna börja hallucinera.

Demokrati och AI: påverkan på beslutsfattande och opinionsbildning

AI och ojämlikhet: risker för ökade klyftor i samhället

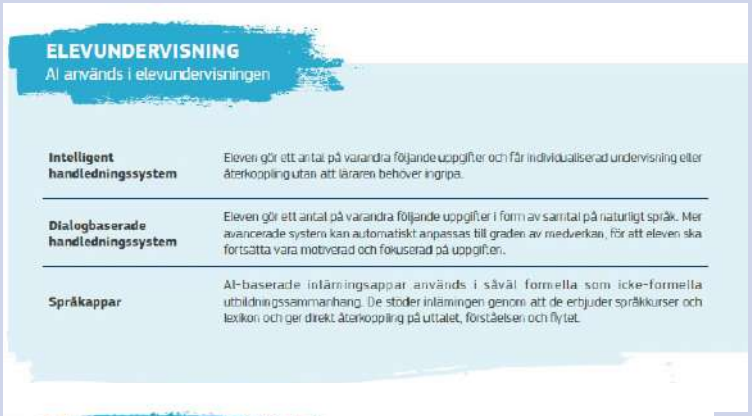
Kulturell påverkan och kreativitet

Att undervisa om AI-etik

https://www.mynewsdesk.com/se/malmo-universitet/pressreleases/ai-kan-loesa-problem-i-skolan-men-aer-ocksaa-ett-potentiellt-hot-3303824

- Metoder för att integrera etiska diskussioner i AI-undervisningen
- Praktiska övningar och case studies för elever
- Källkritik och kritiskt tänkande kring AI-information

Användning av AI och data i utbildningen

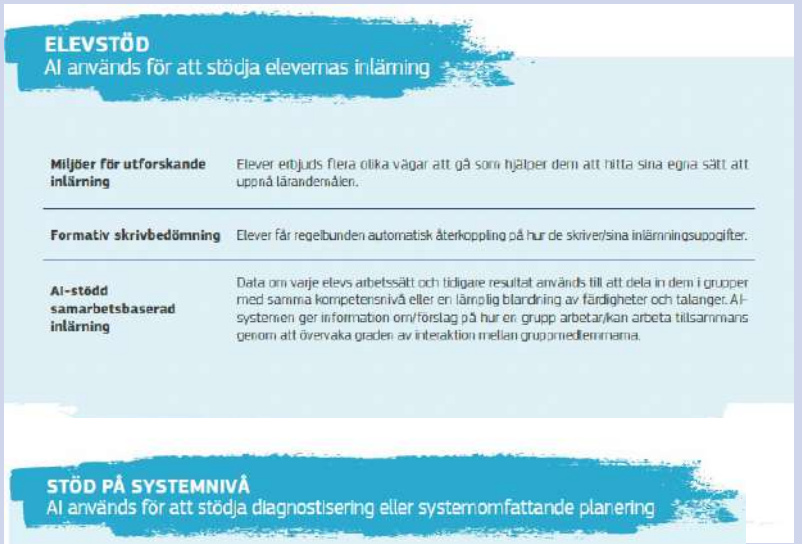


ELEVUNDERVISNING
AI används i elevundervisningen

Intelligent handledningssystem	Eleven gör ett antal på varandra följande uppgifter och får individualiserad undervisning eller återkoppling utan att läraren behöver ingripa.
Dialogbaserade handledningssystem	Eleven gör ett antal på varandra följande uppgifter i form av samtal på naturligt språk. Mer avancerade system kan automatiskt anpassas till graden av medverkan, för att eleven ska fortsätta vara motiverad och fokuserad på uppgiften.
Språkappar	AI-baserade inlärningsappar används i såväl formella som icke-formella utbildningssammanhang. De stöder inlärningen genom att de erbjuder språkkurser och lexikon och ger direkt återkoppling på uttalet, förståelsen och flytet.

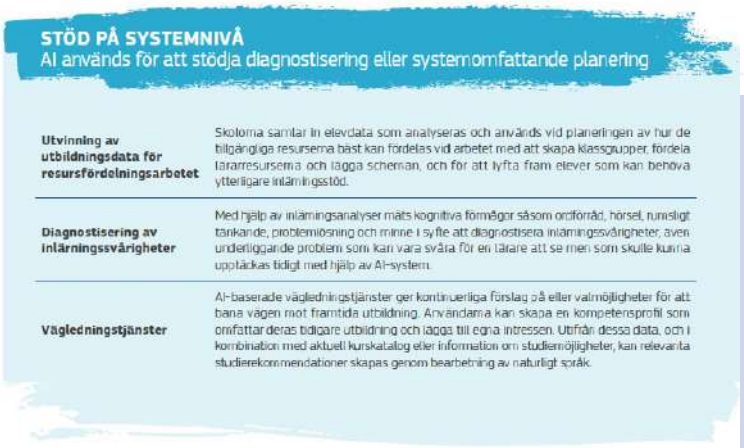
LÄRARSTÖD
AI används för att stödja läraren

Summativ skrivbedömning, poängsättning på uppsatser	AI används till att automatiskt utvärdera och betygsätta elevers skriftliga arbete. AI- och maskininlärningstekniker identifierar egenskaper såsom ordanvändning, grammatik och meningsbyggnad för att sätta betyg och ge återkoppling.
Övervakning av elevforum	Vissa ord i inlägg på elevforum utlöser automatisk återkoppling. Diskussionsanalyser ger inblick i elevernas aktivitet på forumet och kan visa vilka elever som kan behöva hjälp eller som inte deltar som förväntat.
AI-lärarassistenter	AI-agenter eller chattbottar svarar på elevernas vanliga frågor och ger enklare instruktioner och anvisningar. Med tiden kommer AI-systemet att kunna erbjuda allt fler svar och alternativ.
Rekommendationer om pedagogiska resurser	AI-rekommendationsmotorer kan rekommendera särskild undervisning eller särskilda resurser utifrån varje elevs preferenser, framsteg och behov.



ELEVSTÖD
AI används för att stödja elevernas inlärning

Miljöer för utforskande inlärning	Elever erbjuds flera olika vägar att gå som hjälper dem att hitta sina egna sätt att uppnå lärandemålen.
Formativ skrivbedömning	Elever får regelbunden automatisk återkoppling på hur de skriver/sina inlärningsuppgifter.
AI-stödd samarbetsbaserad inlärning	Data om varje elevs arbetssätt och tidigare resultat används till att dela in dem i grupper med samma kompetensnivå eller en lämplig blandning av färdigheter och talanger. AI-systemen ger information om/förslag på hur en grupp arbetar/kan arbeta tillsammans genom att övervaka graden av interaktion mellan gruppmedlemmarna.



STÖD PÅ SYSTEMNIVÅ
AI används för att stödja diagnostisering eller systemomfattande planering

Utvinning av utbildningsdata för resursfördelningsarbetet	Skolorna samlar in elevdata som analyseras och används vid planeringen av hur de tillgängliga resurserna bäst kan fördelas vid arbetet med att skapa klassgrupper, fördela lärarresurserna och lägga scheman, och för att lyfta fram elever som kan behöva ytterligare inlärningsstöd.
Diagnostisering av inlärningssvårigheter	Med hjälp av inlärningsanalyser mäts kognitiva förmågor såsom ordförråd, hörsel, rumsligt tänkande, problemlösning och minne i syfte att diagnostisera inlärningssvårigheter, även underliggande problem som kan vara svåra för en lärare att se men som skulle kunna upptäckas tidigt med hjälp av AI-system.
Vägledningstjänster	AI-baserade vägledningstjänster ger kontinuerliga förslag på eller valmöjligheter för att bana vägen mot framtida utbildning. Användarna kan skapa en kompetensprofil som omfattar deras tidigare utbildning och lägga till egna intressen. Utifrån dessa data, och i kombination med aktuell kurskatalog eller information om studiemöjligheter, kan relevanta studierekommendationer skapas genom bearbetning av naturligt språk.

Etiska riktlinjer för lärare avseende användningen av artificiell intelligens (AI) och data vid undervisning och inlärning https://tinyurl.com/yjfhefbb

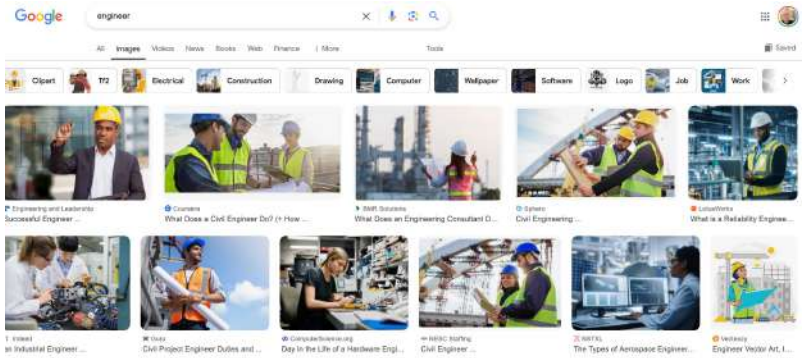
Bias exempel

Ibland tror man att LLMs
är skyldiga till att deras resultat
är partiska och fördomsfulla.
Men de är tränade på stora mängder
historiskt data som visar brist på
jämlighet i olika sammanhang.
Bias kan komma från både
- data
- algoritmer

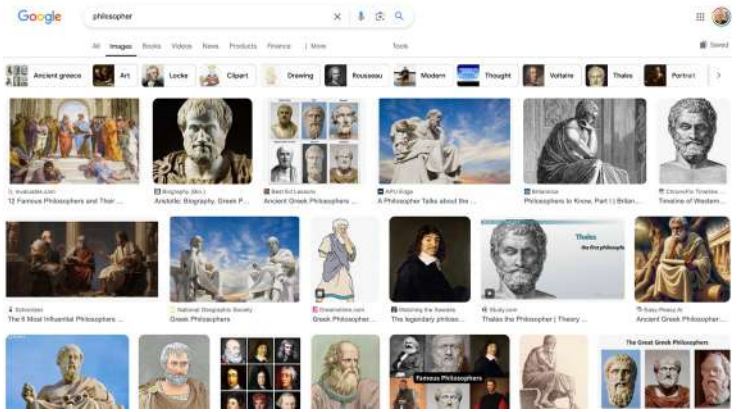
Google har nyligen försökt aktivt
medverka diskriminering pga ras och
kön.

Prompt: En gammal kvinnlig filosof studerar intelligens
i levande organismer - celler, djur, växter

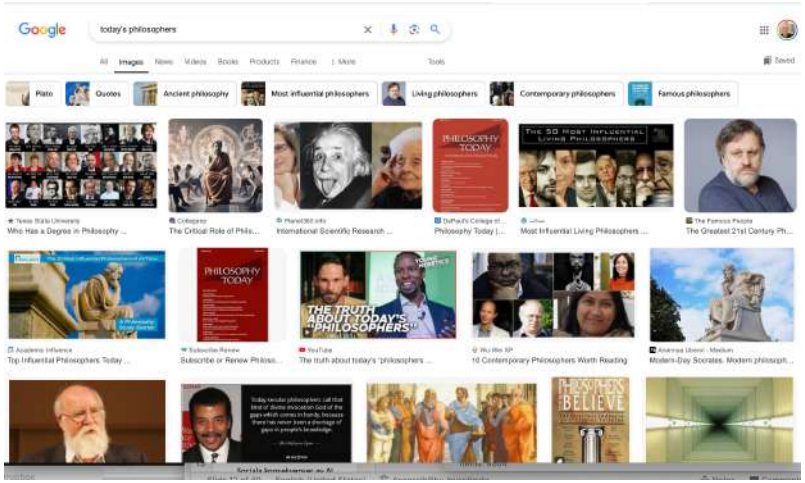
Bias exempel



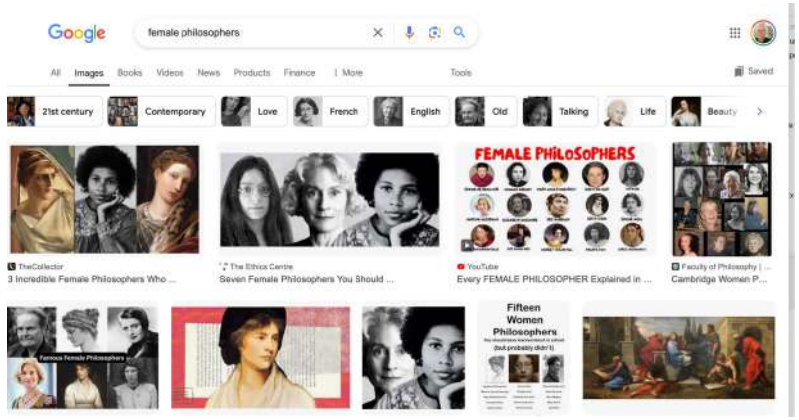
Google sök på “ingenjör”: både manliga och kvinnliga bilder



Google sök på “filosof”: enbart manliga bilder

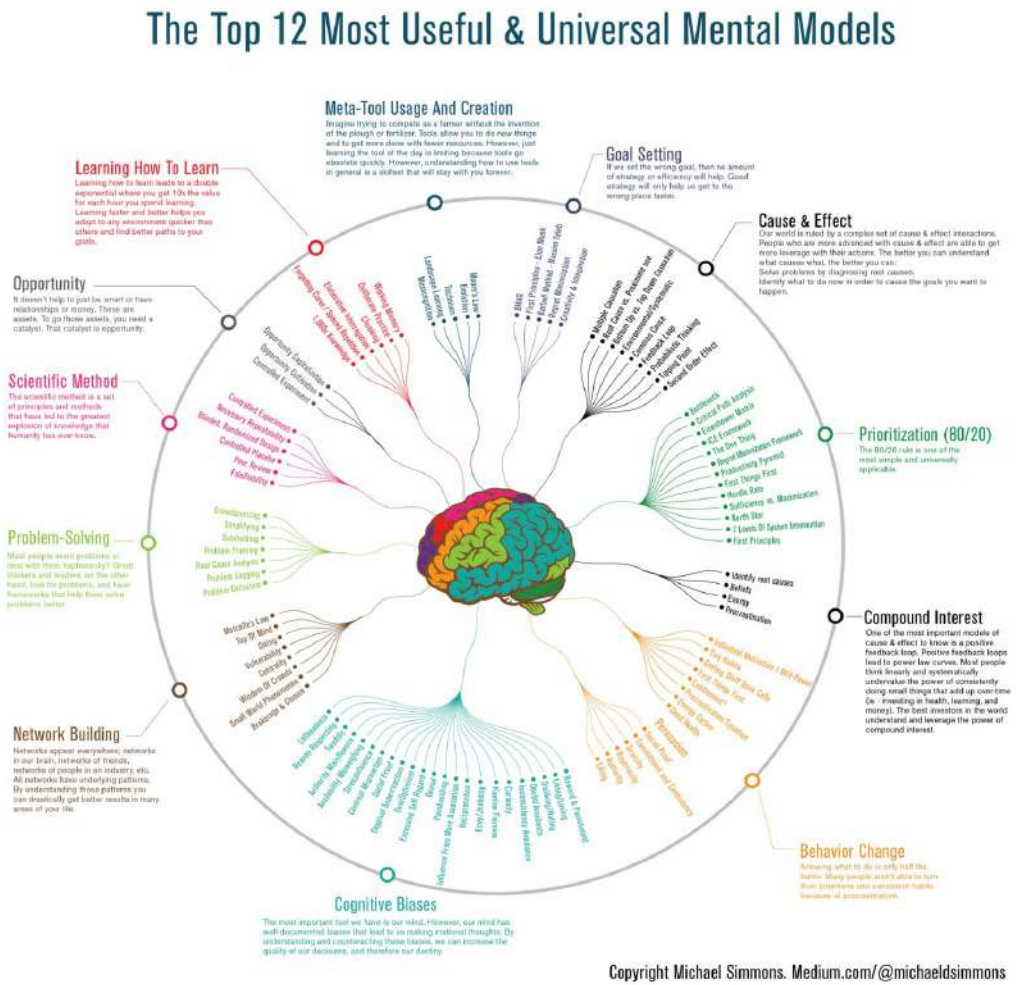


Google sök på “today's philosophers”: även kvinnor förekommer



Google sök på “female philosophers”- de finns!

HUMAN COGNITIVE BIASES

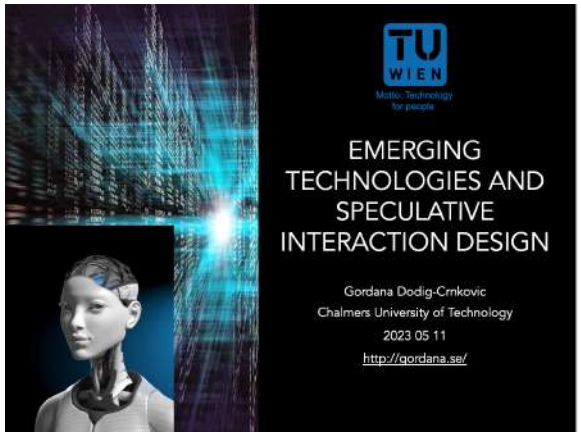


Sociala konsekvenser av AI

https://www.youtube.com/watch?v=Mccpq8fpUI8

- Arbetsmarknadens förändring och automatisering
- AI:s påverkan på utbildning och lärande
- Sociala medier och informationsspridning
- AI i vården och etiska dilemman

Framtidsperspektiv



https://tinyurl.com/3s3784dc

https://tinyurl.com/3hbyfts4

Potentiella framtida etiska utmaningar med AI

Vikten av tvärvetenskapligt samarbete inom AI-etik

Elevernas roll i att forma en etisk AI-framtid

https://www.youtube.com/watch?v=Ctuhh8VqtfI

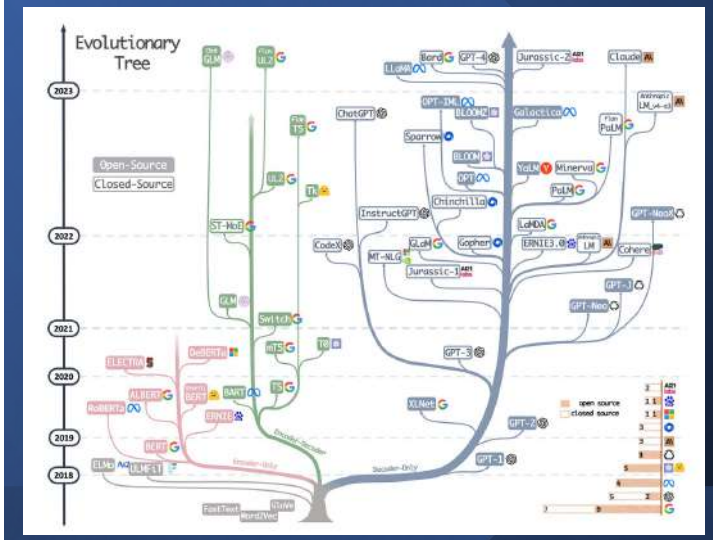
Fei Fei Li, Professor, Stanford University on the History and Future of AI at Data + AI Summit 2024

AI SPRING STARTING IN NOVEMBER 2022

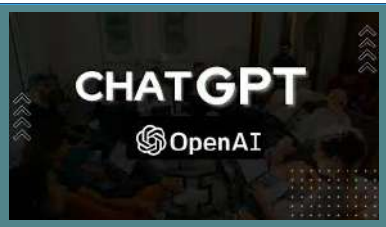
ChatGPT was launched on November 30, 2022,

by San Francisco–based OpenAI

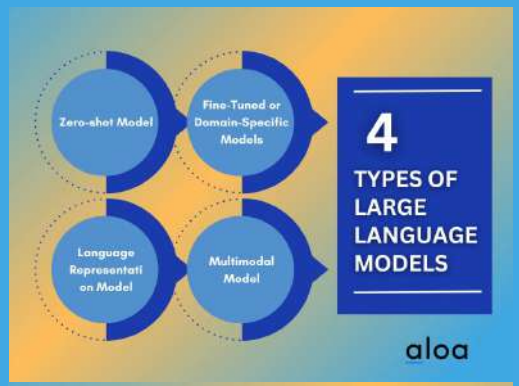
(the creator of the initial GPT series of large language models; DALL. E 2, a diffusion model used to generate images; and Whisper, a speech transcription model).

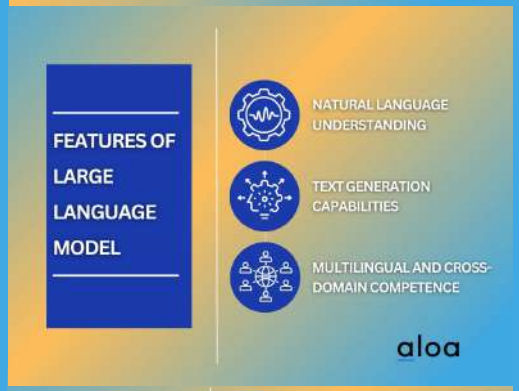


The evolutionary tree of modern LLMs
https://arxiv.org/abs/2304.13712



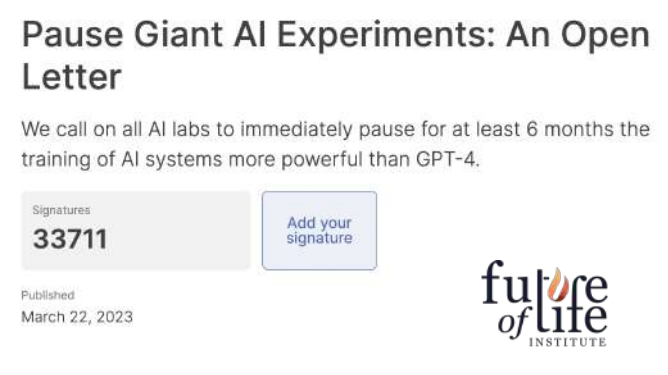
https://backlinko.com/chatgpt-alternatives
https://mobisoftinfotech.com/resources/blog/chatgpt-alternatives/





https://aloa.co/blog/what-is-a-large-language-model-a-beginners-guide

Responses to the dramatic development of AI
Examples of collective action



Signatories include: Yoshua Bengio, Stuart Russell, Gary Marcus, Emad Mostaque, Elon Musk, Tristan Harris, Steve Wozniak and Yuval Noah Harari, Max Tegmark

Geoffrey Hinton and Yoshua Bengio warned in May 2023:

"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war," The letter published by nonprofit organization Center for AI Safety.

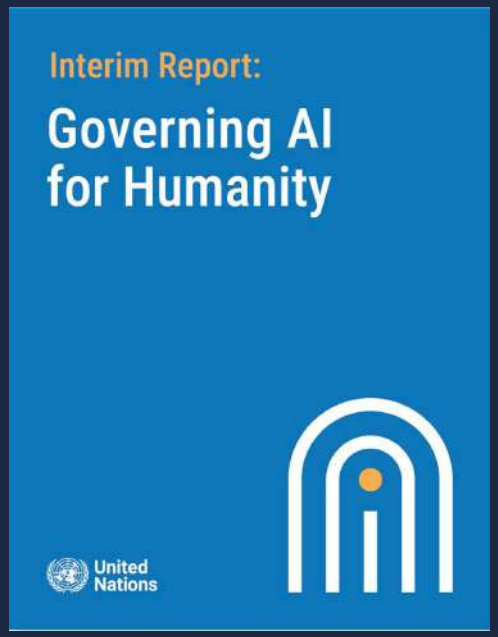
Other signatories include researchers from the Vector Institute and Mila, as well as professors from universities across Canada. Open AI CEO Sam Altman, Microsoft CTO Kevin Scott, etc.

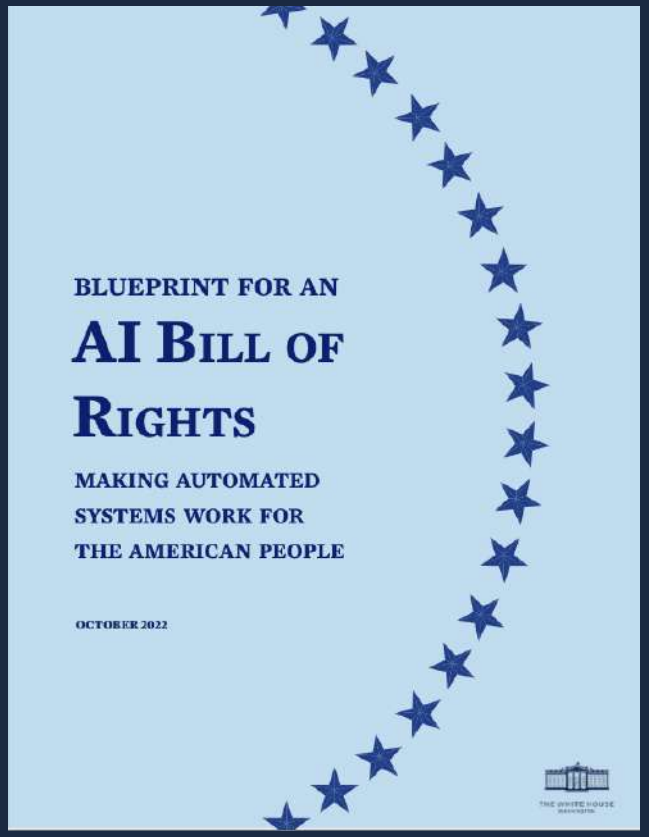
Academics, CEOs sign on in support of AI regulation and Bill C-27 as Canadian companies race to adopt the technology

Since 2023, work on AI regulation

United Nations report (2023)
"Governing AI for Humanity"

https://w.un.org.techenvoy/files/ai_advisory_body_interim_report.pdf





https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

The US AI Bill of Rights outlines principles, including that people have a **right to control how their data is used** and **to not be discriminated against by unfair algorithms.**

It is a white paper, which does not have the force of law. It's primarily aimed at **the federal government** and could influence **which technologies government agencies acquire**, or help parents, workers, policymakers, and designers **ask tough questions about artificial intelligence systems**.

However, **it can't constrain large tech companies**, which arguably play a bigger role in shaping future applications of AI.

EU's "AI Act" (2024)

The world's first AI legislation

AI Act, European Commission. Shaping Europe's digital future

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Source: ISACA

The European Parliament granted final approval of the EU Artificial Intelligence Act on March 13, 2024, by a vote of 523 for passage, 46 against, and 49 abstaining. The Act faces a final step – approval by EU member states – as its provisions gradually take effect.

Can we trust intelligent systems?

Despite anecdotes from people who believed the GPT program was conscious and might persuade humans to behave irrationally and dangerously,

GPT of today has nothing to be conscious with. The reason is similar to the fact that the GPT program does not breathe - it has nothing to breathe with. Looking behind the scenes we can see that present AI is not conscious.

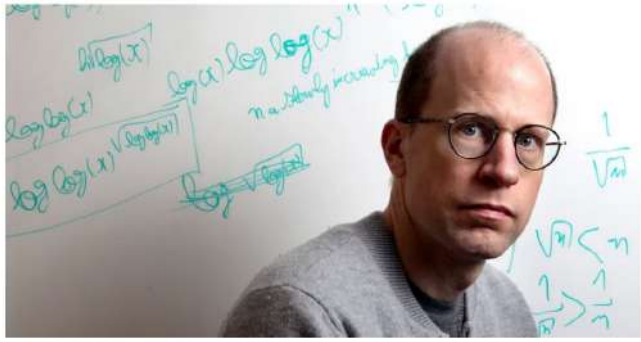
Global reactions to the emergence of GPT programs show how we humans are dependent on and deeply fascinated by language.

Before GPT, there was no entity capable of reasonable dialogue, and not having functioning consciousness.

How is that possible for GPT?

However, it does not mean that in the future a machine cannot be designed to be conscious.

https://www.wired.com/story/nick-bostrom-fear-ai-fix-everything/?bxid=5cec24ecfc942d3ada053185&cndid=55112229&esrc=bounceX&source=Email_0_EDT_WIR_NEWSLETTER_0_DAILY_ZZ&utm_brand=wired&utm_campaign=aud-dev&utm_content=WIR_Daily_050524&utm_mailing=WIR_Daily_050524&utm_medium=email&utm_source=nl&utm_term=WIR_Daily_Active



FAST FORWARD | 5-MINUTE READ

Nick Bostrom Made the World Fear AI. Now He Asks: What If It Fixes Everything?

BY WILL KNIGHT

Philosopher Nick Bostrom popularized the idea that superintelligent AI could erase humanity. His new book imagines a world in which algorithms have solved every problem.

FEAR OF ARTIFICIAL INTELLIGENCE?
NLP, ML AND LLMS BASED DISCOVERY OF AI-PHOBIA AND FEAR SENTIMENT PROPAGATION BY AI NEWS

WORKING PAPER (PREPRINT) – RELEASED FOR RAISE-24 ON 3/9/2024

Jim Samuel
Rutgers University
jim.samuel@rutgers.edu

Tanya Khanna
Rutgers University

Srinivasaraghavan Sundar
Rutgers University

ABSTRACT

Confusion, fear and mixed sentiments prevail in the minds of people towards what is arguably one of the most important of dynamics of modern human society: Artificial Intelligence (AI). This study aims to explore the contributions of news media towards this phenomenon - we analyze nearly seventy thousand recent news headlines on AI, using natural language processing (NLP) informatics methods, machine learning (ML) and large language models (LLMs) to draw insights and discover dominant themes. Our theoretical framework was derived from extant literature which posits the power of fear producing articles and news headlines which produce significant impacts on public behavior even when available in small quantities. We applied extensive textual informatics methods using word and phrase frequency analytics, sentiment analysis and human experts based thematic analysis to discover insights on AI phobia inducing news headlines.

Our rigorous analysis of nearly seventy thousand headlines using multiple validation methods in NLP (exploratory informatics including BERT, Llama 2 and Mistral[1; 2] based topic identification), ML (supervised informatics) and LLMs (neural nets for sentiment classification, with BERT, Llama 2 and Mistral) demonstrates the presence of an unreasonable level of emotional negativity and fear inducing verbiage in AI news headlines. The framing of AI as being dangerous or as being an existential threat to humanity can have a profound impact on public perception, and the resulting AI phobia and confusion in public perceptions are inherently detrimental to the science of AI. Furthermore, this can also impact AI policy and regulations, and harm society. We conclude with a discussion deducing implications for society and make recommendations for education and policies that could support human identity and dignity.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4755964

“Domestication of ignorant entities”

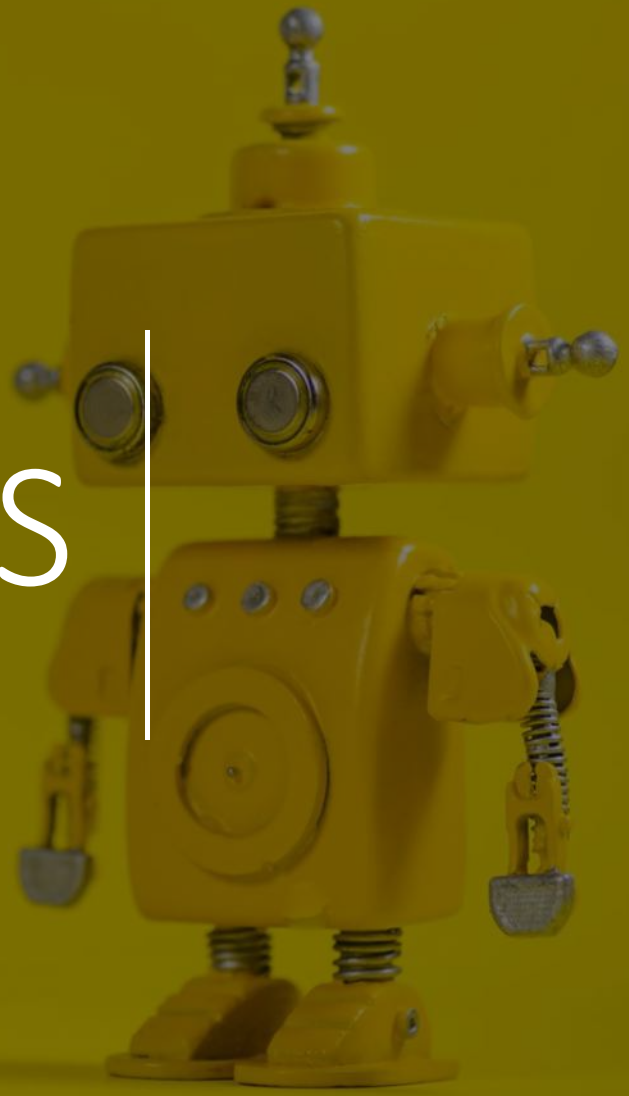
“Eco-cognitive computationalism considers computation in context, following some of the main tenets advanced by the recent cognitive science views on embodied, situated, and distributed cognition. ”

“Through eco-cognitive computationalism we can clearly acknowledge that the concept of computation changes, depending on historical and contextual causes, and we can build an epistemological view that illustrates the “emergence” of new kinds of computations, such as the one regarding morphological computation. This new perspective shows how the computational domestication of ignorant entities can originate new unconventional cognitive embodiments.”

Lorenzo Magnani (2021) Computational domestication of ignorant entities Synthese 198(11)
DOI: 10.1007/s11229-020-02530-5

ROBOTS

Humanoid robots
Education robots
Consumer robots
Research robots
Medical robots
Nano robots
Disaster response robots
Industrial robots
Aerospace robots
Underwater robots
Aerospace robots
Military and Security robots
Telepresence robots
Drones
Autonomous cars



https://robots.ieee.org/

Autonomous Cars

https://robots.ieee.org/

Boss

Google Self-Driving Car

nuTonomy

Stanley

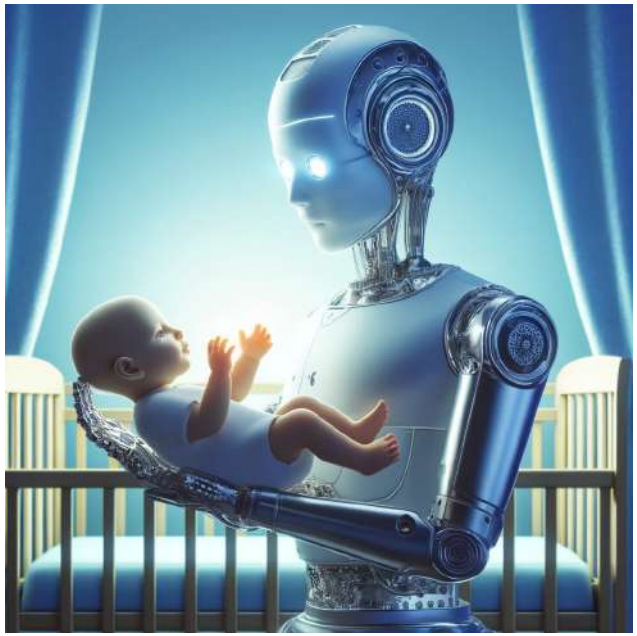
Waymo

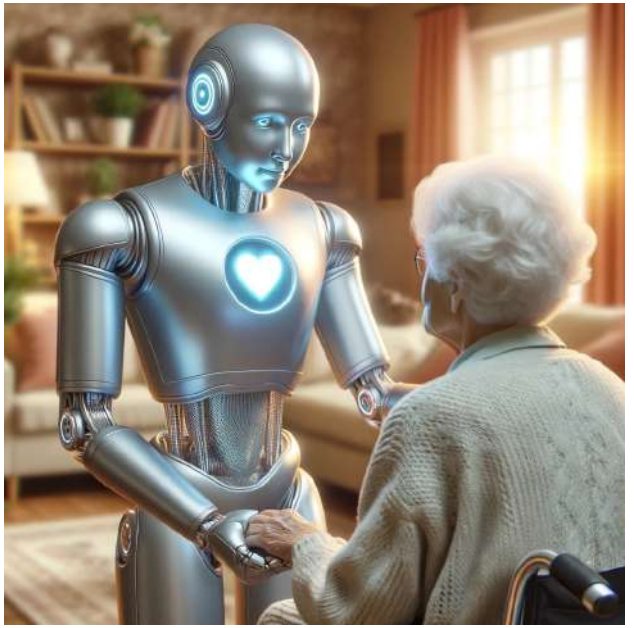
A flame - throwing robot

https://www.youtube.com/watch?v=U83BfU1phCw

https://www.youtube.com/watch?v=b5dE3vxWP9E

Intelligent, ethical robot according to GPT4 (Dall-E)





Delegating responsibilities to intelligent autonomous systems: challenges and benefits

Gordana Dodig-Crnkovic, Gianfranco Basti, and Tobias Holstein

United Nations report "Governing AI for Humanity" and EU's "AI Act" emphasize the human role in ethical AI development, advocating for inclusive governance and continuous ethical oversight of socio-technological systems. We explore the concept of distributed responsibility in a network of agents, drawing on perspectives that distinguish between human ethical deliberation and machine responsiveness where AI is seen as a part of a larger interconnected system with shared responsibilities.

It is important to acknowledge the limitations of human judgment and actively work towards mitigating its consequences through careful design, the use of diverse competencies, continuous oversight, and constant systemic learning.

The discussion extends to the machine ethics approach, which integrates ethical principles into AI design, aiming for consistency, scalability, and alignment with human values. We argue for a multifaceted strategy that includes continuous learning, ethical education, and societal engagement to ensure the development of responsible AI. We identify the limitations of human judgment and the necessity for meticulous design and oversight to navigate the ethical landscape of AI integration into society.

THE ROLE OF HUMANS: TIME PERSPECTIVE

Short-term perspective
We, humans, decide

Middle-term perspective
AGI & We co-decide

Long-term perspective

Superintelligence? Who decides?

Constantly. Learning from experience. Feedback on development & design

ASSIGNMENT OF RESPONSIBILITY: WHO DECIDES?

Levels of AI

- ANI (Narrow AI)
- AGI (Artificial General Intelligence)
- ASI (Artificial Super Intelligence)

Stakeholders

- Politicians
- Legislators
- Businesses
- Requirements engineers
- Designers, Developers
- Programmers
- Deployment engineers, testers
- Maintenance engineers

https://tinyurl.com/pjbdyn95 Global AI governance: barriers and pathways forward- Huw Roberts, Emmie Hine, Mariarosaria Taddeo, Luciano Floridi

WHAT CAN WE LEARN FROM AUTONOMOUS CARS ABOUT ETHICS ASPECTS OF OTHER ROBOTS?

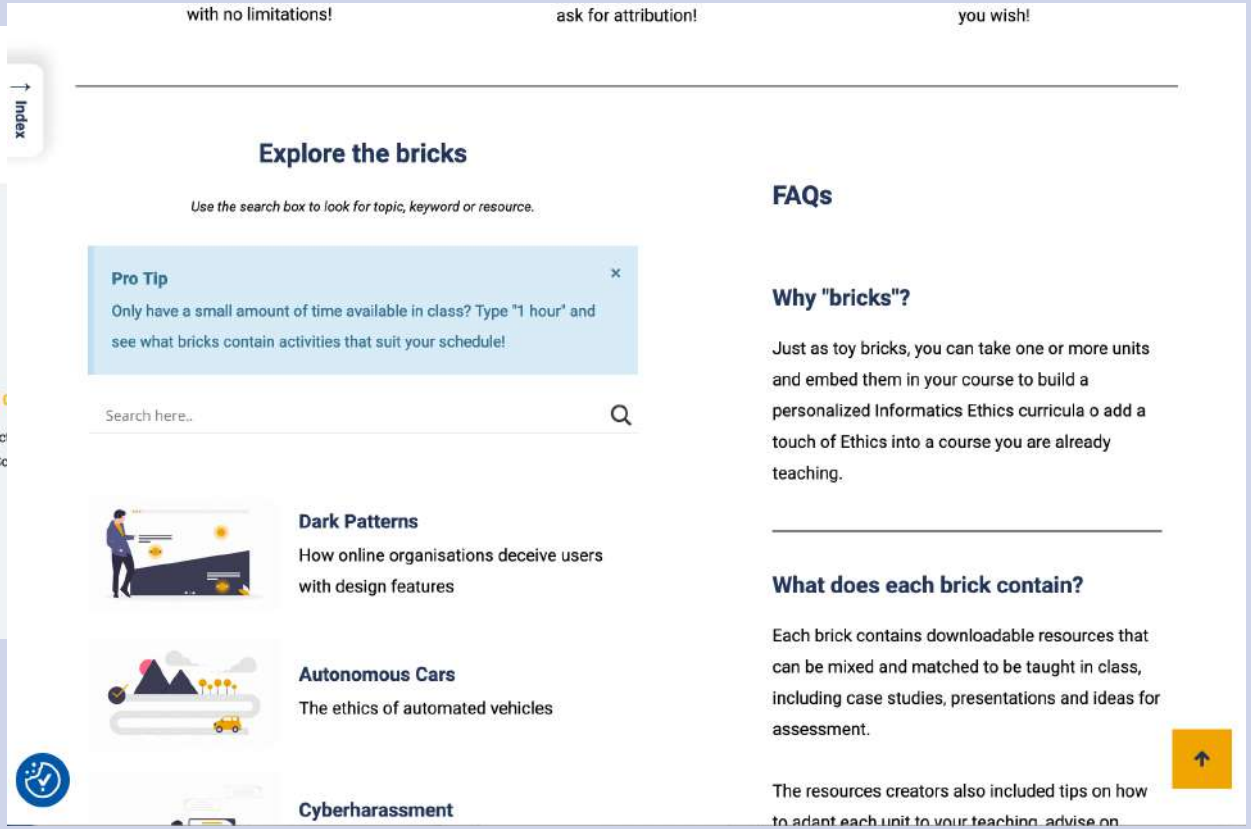
Autonomous Cars

Based on:

Holstein, T., Dodig-Crnkovic, G., & Pelliccione, P. (2021). Steps Towards Real-world Ethics for Self-driving Cars: Beyond the Trolley Problem. In Steven John Thompson (Ed.), Machine Law, Ethics, and Morality in the Age of Artificial Intelligence. IGI Global

Picture: https://www.aarete.com/insights/what-is-the-business-case-for-autonomous-vehicles-in-the-supply-chain/

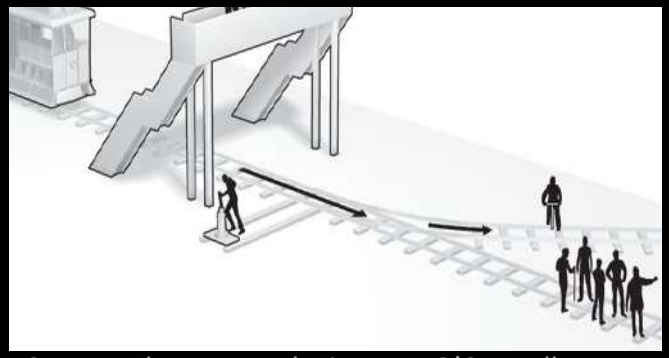
ETHICS4EU EDUCATIONAL BRICK ON AUTONOMOUS CARS



https://ascnet.ie/ethics4eu-website/

https://drive.google.com/file/d/1kEn8yfLes-a7LjkFLioEeaC-kFN_nNoX/view?pli=1

THE "TROLLEY PROBLEM"- THE PSEUDO-PROBLEM DOMINATING FOR LONG TIME THE DEBATE ABOUT AUTONOMOUS CARS



Source: The New York Times; F. O'Connell

Current discussions about self-driving cars repeatedly take form of decision-making problem borrowed from philosophy

THE TROLLEY PROBLEM: Whom will the self-driving car kill when it has to decide?
Example:
Bonnefon, Shariff och Rahwan The social dilemma of autonomous vehicles Download The social dilemma of autonomous vehicles

Ethical thought experiment defined by philosopher Philippa Foot in "The Problem of Abortion and the Doctrine of the Double Effect," pp. 5-15, *Oxford Review,* 5, (1967). Focus on the difference between responsibility of acting vs. non-acting.

Many different variants, such as the use of personas to include an emotional perspective. But there is always a single decision to make: Whom to kill? The Trolley Problem is Unsolvable by Construction

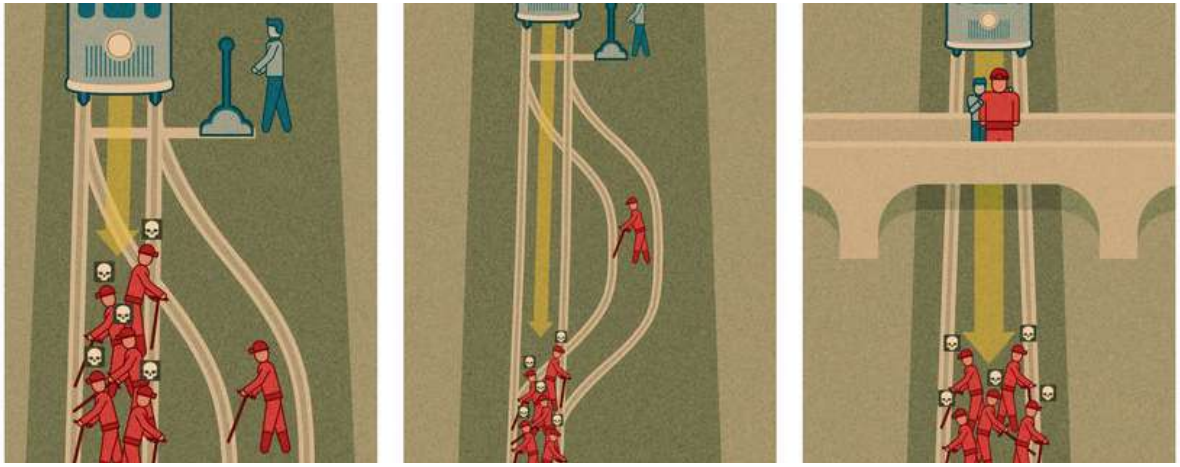
The Original Trolley Problem –
The Doctrine of Double Effect in Medicine, in life-and-death situations of abortion & euthanasia

The **Doctrine of Double Effect**, DDE: It is morally acceptable to perform an action that has both a positive effect and a harmful one, provided that:
1.The harmful effect is not intended (even if it is foreseen).
2.The harmful effect is not the means by which the good effect is achieved.
3.There is a proportionate reason for allowing the harmful effect.

Philippa Foot's 1967 Paper
In her paper *The Problem of Abortion and the Doctrine of Double Effect*, Foot explores how the DDE applies to complex moral situations, particularly focusing on issues like **abortion** and **euthanasia**. She critically examines whether the doctrine provides a sufficient framework for making moral judgments in these cases.
Foot discusses the differences between ***killing*** *and* ***letting die*** *as well as* ***negative duty*** *(the duty not to harm others) and our* ***positive duty*** *(the duty to help others).*

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I. The Moral Machine experiment. Nature. 2018 Nov;563(7729):59-64. doi: 10.1038/s41586-018-0637-6. Epub 2018 Oct 24. PMID: 30356211.

The Trolley Problem for Runaway Trolley



https://www.mpg.de/14386104/trolley-dilemma-interntional

By using the Trolley Problem, Philippa Foot not only questioned the robustness of the Doctrine of Double Effect but also opened the door to a broader philosophical conversation about how we make moral judgments in life-and-death situations.

Applying the Trolley Problem to autonomous cars makes no sense. We should first solve the problem for human drivers. (unsolvable)
The c ar will never make life-and-death **decisions**. Our role as designers and engineers is to prevent accidents, not to cho o se whom to kill.

CORRESPONDENCE | 05 March 2019

'Moral machine' experiment is no basis for policymaking

By Barry Dewitt, Baruch Fischhoff & Nils-Eric Sahlin

The 'moral machine' experiment for autonomous vehicles devised by Edmond Awad and colleagues is not a sound starting place for incorporating public concerns into policymaking (*Nature* **563**, 59–64; 2018).

The experiment presents participants with stylized moral dilemmas that are intended to resemble choices facing designers and regulators. For example, participants must choose between a crash that kills three elderly pedestrians and one that kills three non-elderly occupants of an autonomous vehicle.

The study would have benefited from a premise common to philosophy and psychology: namely, that stylized dilemmas are a means rather than an end. They are meant to pose questions rather than answer them, and to inform public discourse rather than attempt to resolve it (B. Fischhoff *Science* **350**, aaa6516; 2015).

Philosophers use stylized tasks to analyse the complex and uncertain situations in which moral choices are actually made. Dilemmas have no meaning outside such discourse. Although survey responses might stimulate enquiry, taking them literally is an antithesis to philosophical practice.

Psychologists use stylized tasks to test individuals' sensitivity to cues that could help them to decide between options. A single representation of a dilemma cannot stand alone, without knowing how participants interpret it, how they respond to alternative wording and how they view the ethics of a society guided by survey responses (see D. Medin *et al. Nature Hum. Behav.* **1**, 0088; 2017).

Nature **567**, 31 (2019)

doi: https://doi.org/10.1038/d41586-019-00766-x

THE WAY WE MAKE DECISIONS

TUANA. COMMUNICATIONS OF THE ACM | DECEMBER 2015 | VOL. 58 | NO. 12

Values

Values serve as a guide to action and knowledge.

They are relevant to all aspects of scientific and engineering practice, including discovery, analysis, and application.

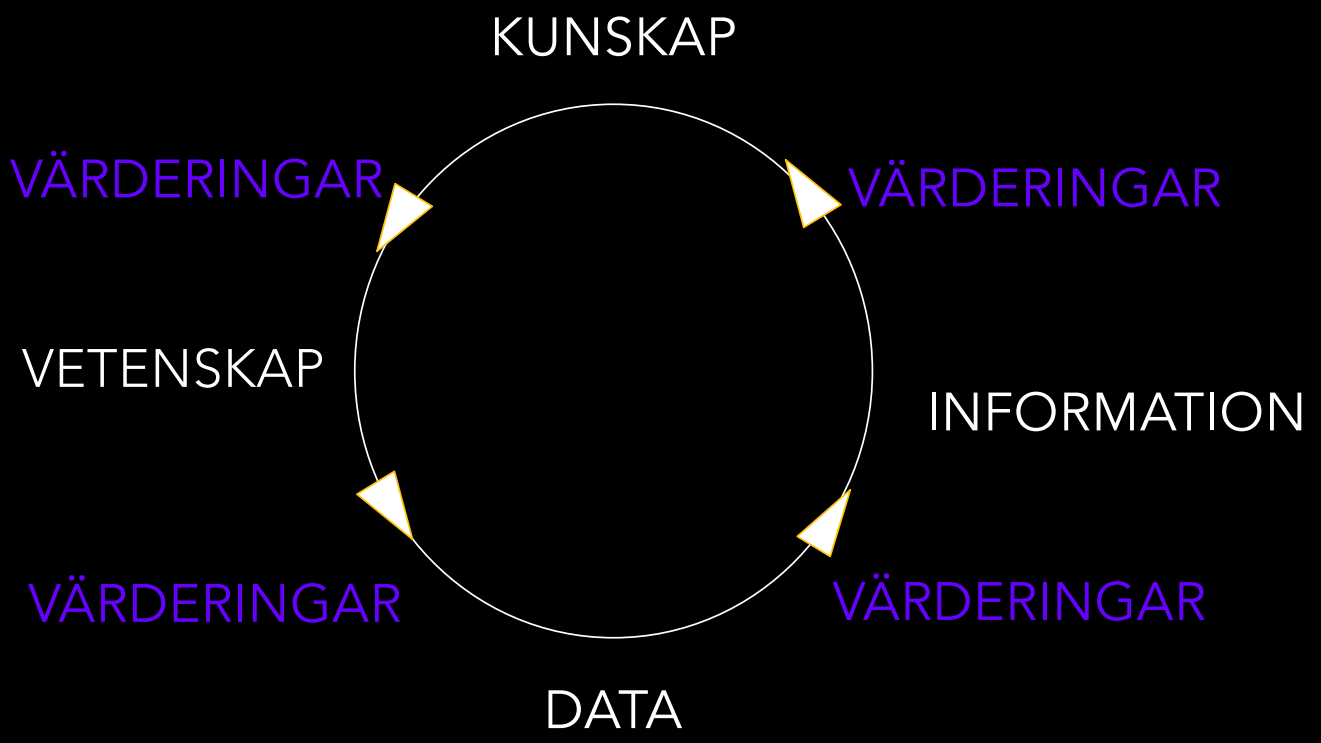
VÄRDERINGAR OCH ETIK I KUNSKAPSPRODUKTION

Baserad på:

Nancy Tuana (2015)
Coupled Ethical-Epistemic Analysis in Teaching Ethics. Critical reflection on value choices.
CACM VOL. 500 NO. 12. Pages 27-29

http://cacm.acm.org/magazines/2015/12/194630-coupled-ethical-epistemic-analysis-in-teaching-ethics/abstract

Värderingar i kunskapsproduktion



Se: https://link.springer.com/article/10.1007/s11023-024-09697-7 **The Inherent Normativity of Concepts**

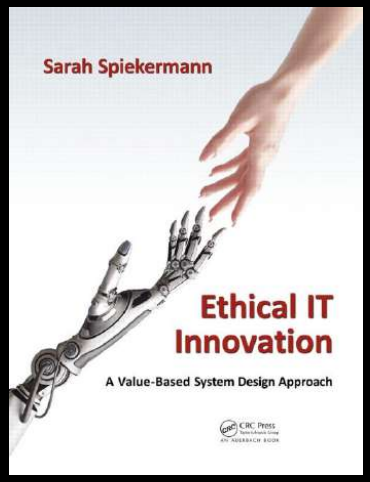
VÄRDERINGAR

Värderingar fungerar som vägledning till handling och kunskap. De är relevanta för alla aspekter av vetenskaplig och teknisk praxis, inklusive upptäckt, analys och tillämpning.

Kognitionsforskare har funnit att värden är en integrerad del av STEM-forskningen (Science, Technology, Engineering, and Mathematics).

Ethical IT Innovation: A Value-Based System Design Approach

Etisk IT-innovation: En värdebaserad systemdesignmetod

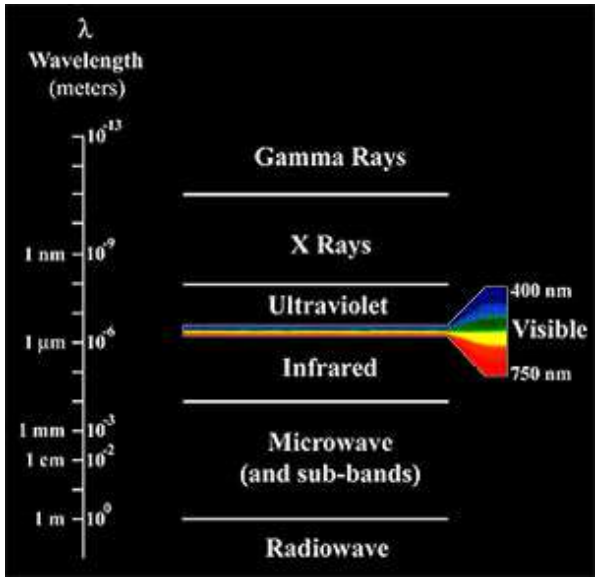


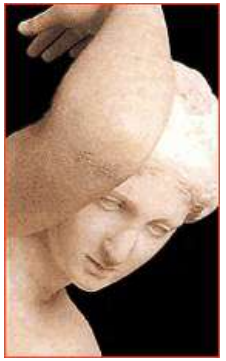
Sarah Spiekermann:

IEEE P7000
The first global standard process for addressing ethical concerns in system design

Förutom Sarah Spiekermann, har vi I Europa en rad världsledande etiker, bl.a. Peter Paul Verbeek, Phillip Brey, Jeroen van den Hoven, Ibo van de Poel, Luciano Floridi, Mariarosaria Taddeo, Vincent Mueller, Raffael Capurro, Virginia Dignum, SO Hansson och manga fler.

Världen kan ses i olika ljus





Tänk om vi kunde se i vilken våglängd som helst av det elektromagnetiska spektrumet, från gammastrålar till radiovågor? Hur skulle världen se ut för oss?

Etiska aspekter i multikriterie-beslutsanalys

Ethical Aspects of Technology in the Multi-Criteria Decision Analysis

Gordana Dodig Crnkovic, Chalmers University of Technology and University of Gothenburg, Sweden
gordana.dodig-crnkovic@chalmers.se

Gaetana Sapienza, ABB Corporate Research and Mälardalen University, Sweden
gaetana.sapienza@se.abb.com

Abstract. In technological systems, decisions are often governed by multi criteria decision analysis (MCDA) techniques that take into account mutually opposing criteria for the system, and it results in ranking of alternatives. MCDA is based on value systems of decision-makers, and ethical deliberation in the process is implicit. We argue that it is necessary to make decision-making in technological systems transparent such that value basis and ethical considerations become explicit and subject for scrutiny of involved stakeholders. As different priorities, value systems and ethical choices result in different technical solutions, such solutions when put in use will promote those intrinsic and implicit values. In a society with ubiquitous technology, value aspects of technology are essential. At present there is no explicit mechanism to expose ethical aspects in these analyses, so they can easily be forgotten. As a support to encourage introduction of transparent value-based deliberation we propose an extended MCDA scheme that explicitly takes into account ethical analysis.

https://tinyurl.com/mre9knw5

Inclusion of Ethical Aspects in Multi-criteria Decision Analysis

Publisher: IEEE | Cite This | PDF

Gaetana Sapienza ; Gordana Dodig-Crnkovic ; Ivica Crnkovic **All Authors**

2 Cites in Papers | 216 Full Text Views

Abstract

Document Sections

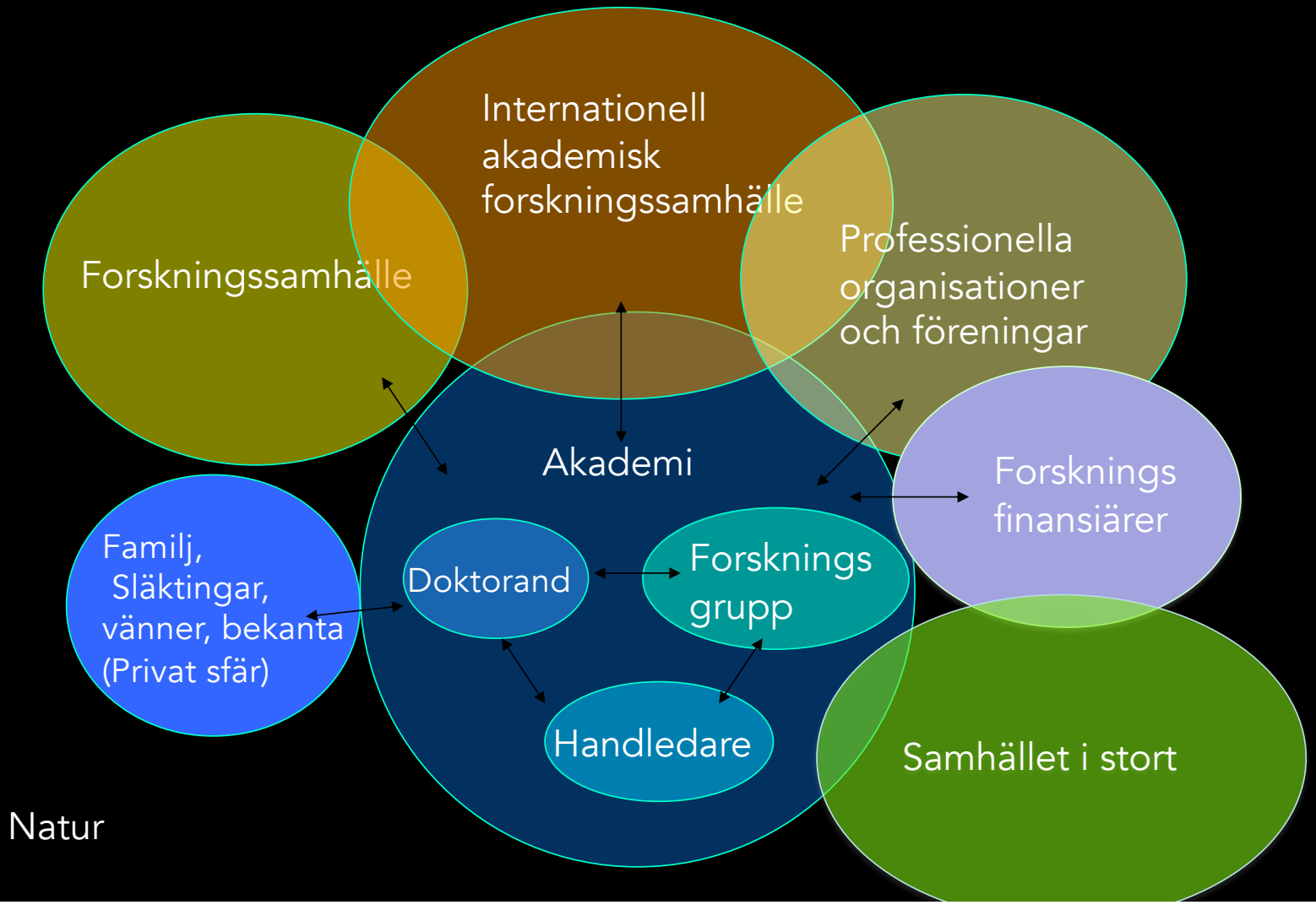
I. Introduction
II. Background
III. The Importance of Ethical Aspects
IV. Explicating Ethical Aspects in MCDA
V. Case Study: HW/SW Partitioning for a Wind Turbine

Abstract:
Decision process is often based on multi-faceted and mutually opposing criteria. In order to provide rigorous techniques for problem structuring and criteria aggregation used for classification and ranking of alternatives, Multiple Criteria Decision Analysis (MCDA) has been used as a method to achieve architectural decisions. Even though it has already been argued in literature that MCDA essentially depends on value systems of decision-makers, it is a question how the decision result reflects a particular criterion, requirement or a particular decision. This is especially true if a criterion is not precisely specified. In this paper we analyse the ethical aspects of MCDA. In our analysis we argue that it is in the long run necessary to make value basis of decision-making and ethical considerations explicit and subject for scrutiny. As a support to encourage introduction of transparent value-based deliberation we propose an extended MCDA scheme that would explicitly take into account ethical analysis. As an illustration, we present an industrial case study for the Software (SW)/Hardware (HW) partitioning of a wind turbine application in which different decisions can be taken, depending on the ethical aspects.

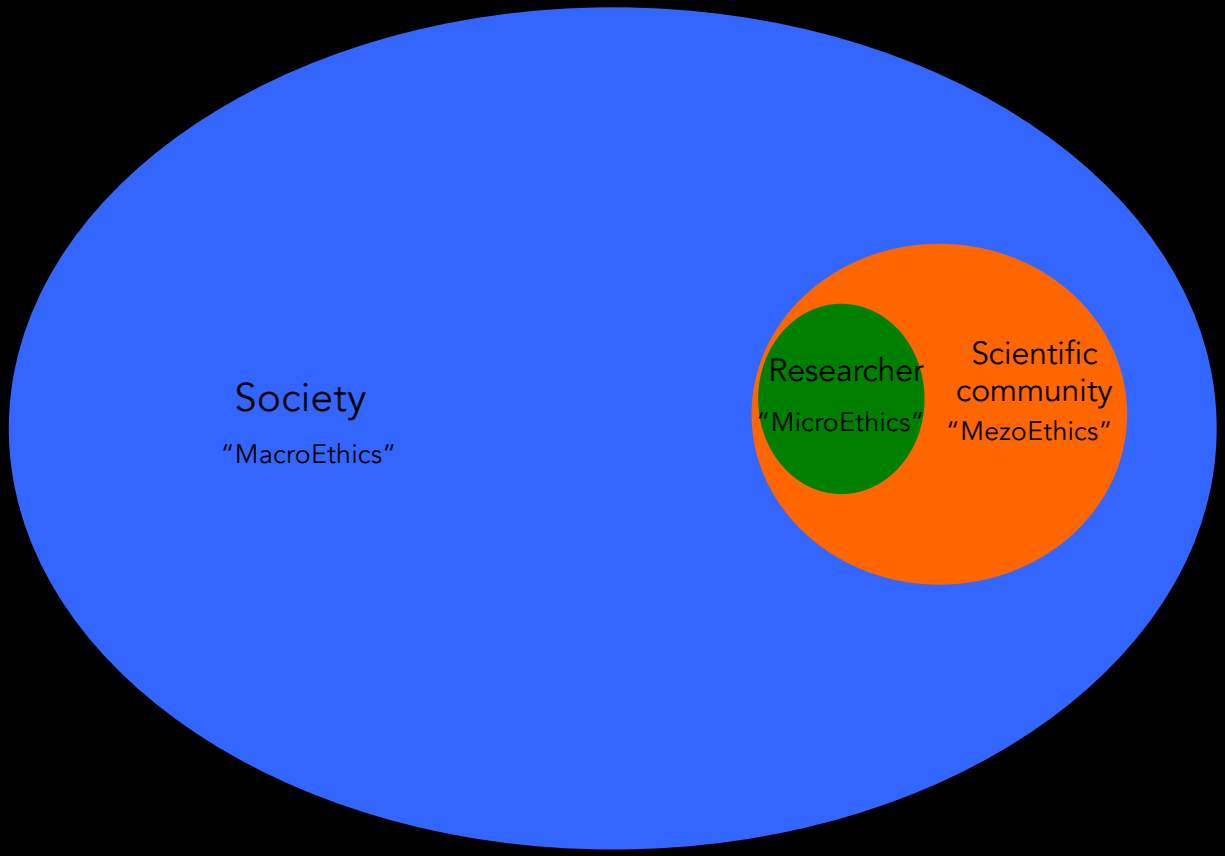
Published in: 2016 1st International Workshop on Decision Making in Software ARCHitecture

https://ieeexplore.ieee.org/document/7496439

AKTÖRER I ETT AKADEMISKT FORSKNINGSPROJEKT



Forskningsetiska domäner



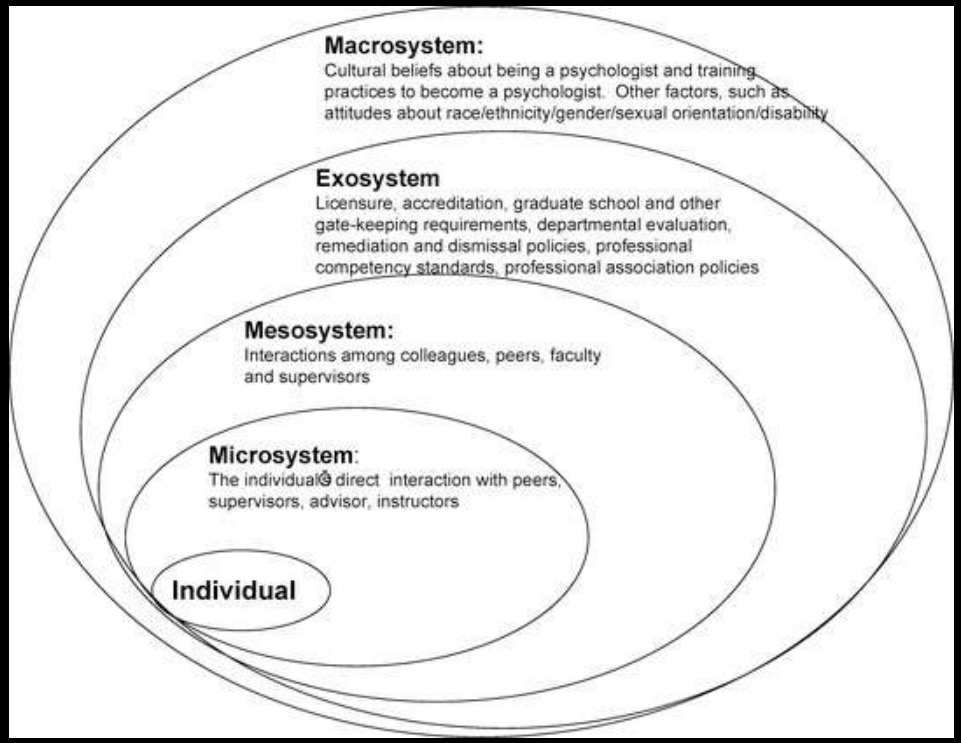
Is it true?
1. Scientific integrity

Is it fair?
2. Collegiality
3. Protection of human subjects
4.Animal welfare
5. Institutional integrity

Is it wise?
6. Social responsibility

Kenneth D. Pimple (2002) "Six Domains of Research Ethics. A Heuristic Framework for the Responsible Conduct of Research". Science and Engineering Ethics 8 , 191-205

Micro – Meso – Exo – Macro Domains



ou will recognize this domain-based view in the analysis of many different types of problems – organization of society, sustainability of cities, ecology, economics, ethical aspects etc.

The human role in ethical AI vs. the role of AI agents

The United Nations report "Governing AI for Humanity" and the EU's "AI Act" emphasize the human role in ethical AI development, advocating for inclusive governance and continuous ethical oversight of socio-technological systems.

We explored the concept of distributed responsibility in a network of agents, drawing on perspectives that distinguish between human ethical deliberation and machine responsiveness.

Autonomous AI is seen as a part of a larger socio-technological interconnected system with shared responsibilities.

More powerful virtual agents
"2023 was the year of being able to chat with an AI. Multiple companies launched something, but the interaction was always you type something in and it types something back," says Stanford's Peter Norvig.
"In 2024, we'll see the ability for agents to get stuff done for you. Make reservations, plan a trip, connect to other services."

https://www.ibm.com/blog/artificial-intelligence-trends/

https://www.youtube.com/watch?v=Boj9eD0Wug8 Mark Zuckerberg & Yuval Noah Harari in Conversation (01:25:00 Totalitarianism & Surveillance Capitalism)

Self-Driving (Autonomous) Cars as Intelligent Robotic System

We take Self-driving cars as an example of emerging technology that is combining advances in several underlying emergent technologies such as electric mobility and artificial intelligence (with connected driving, intelligent cities, intelligent infrastructure, etc.)

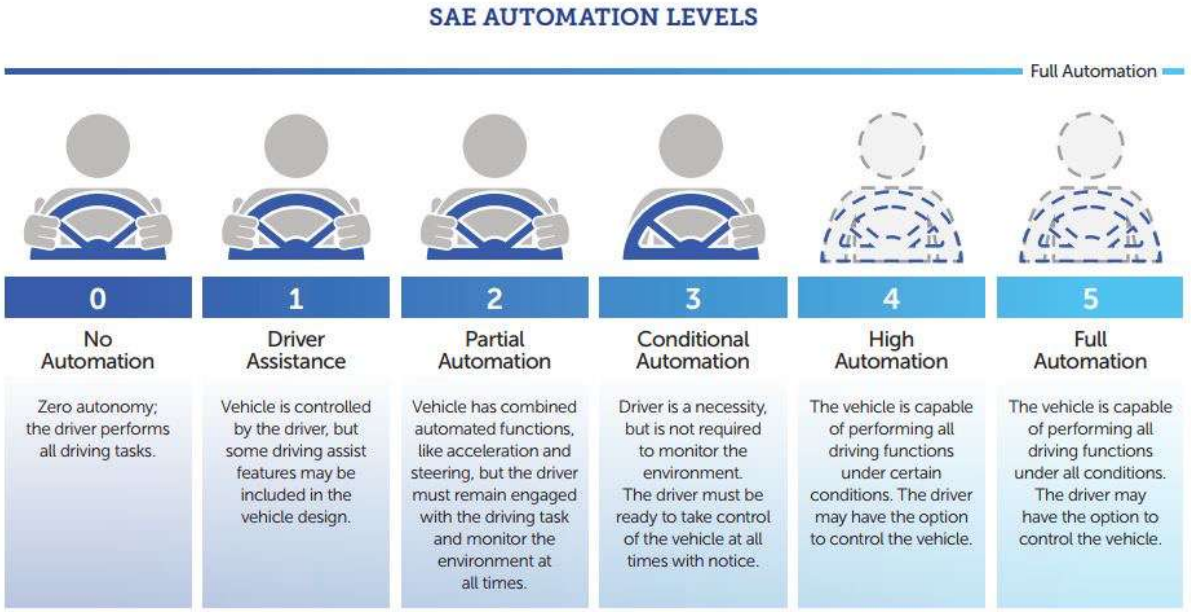
Technology emerges not in vaccuum but in its social context that today is global technosocial environment

How can we contribute in different roles as stakeholders to the development of good society with help of new powerful technologies. Who are the main actors/stakeholders and how do they affect the development? Autonompus cars have been studied a lot and we can learn from the development so far.

AUTONOMOUS CARS DEVELOPMENT

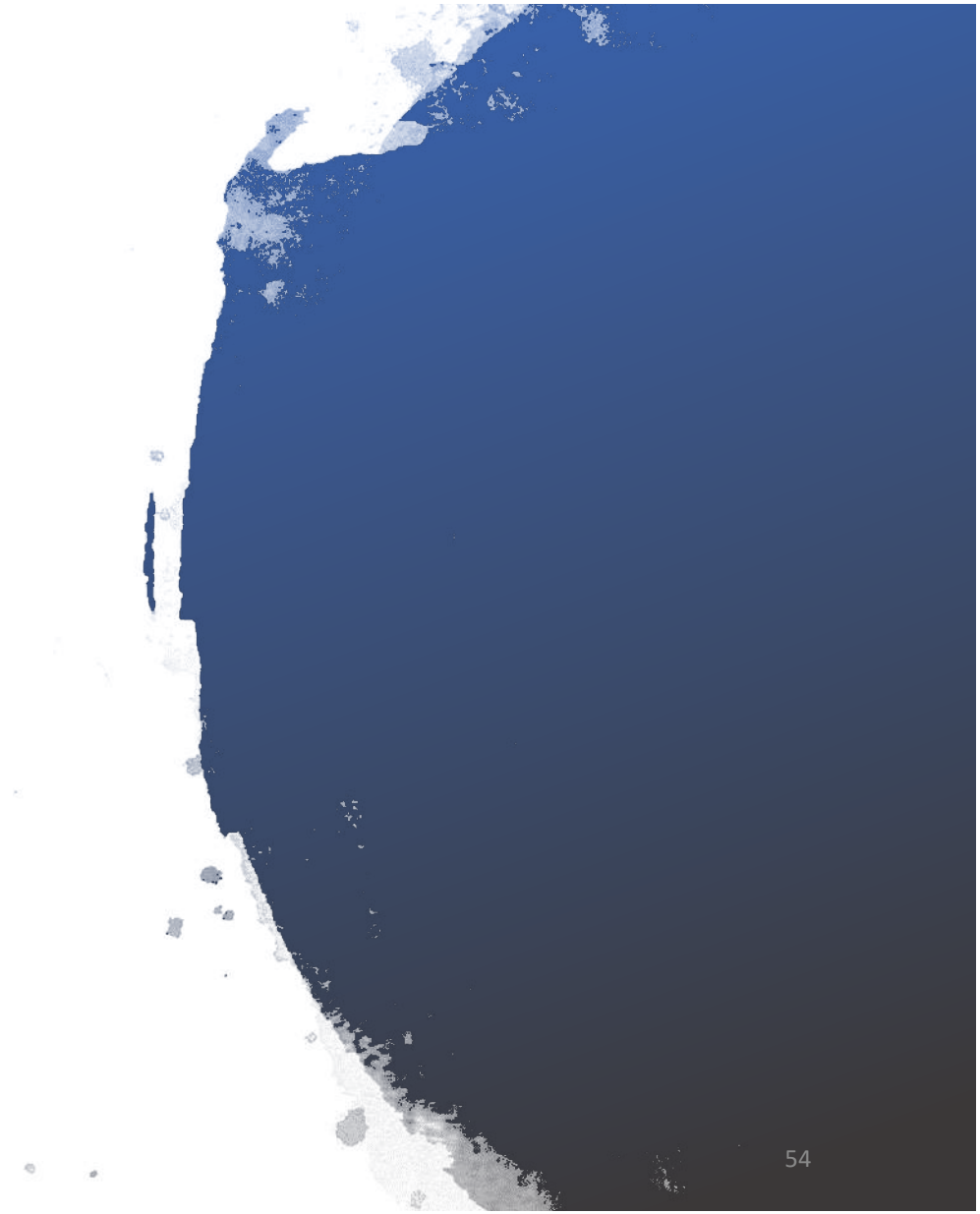
LEVELS OF AUTOMATION

Equivalent for Intelligent Robotics Needed!

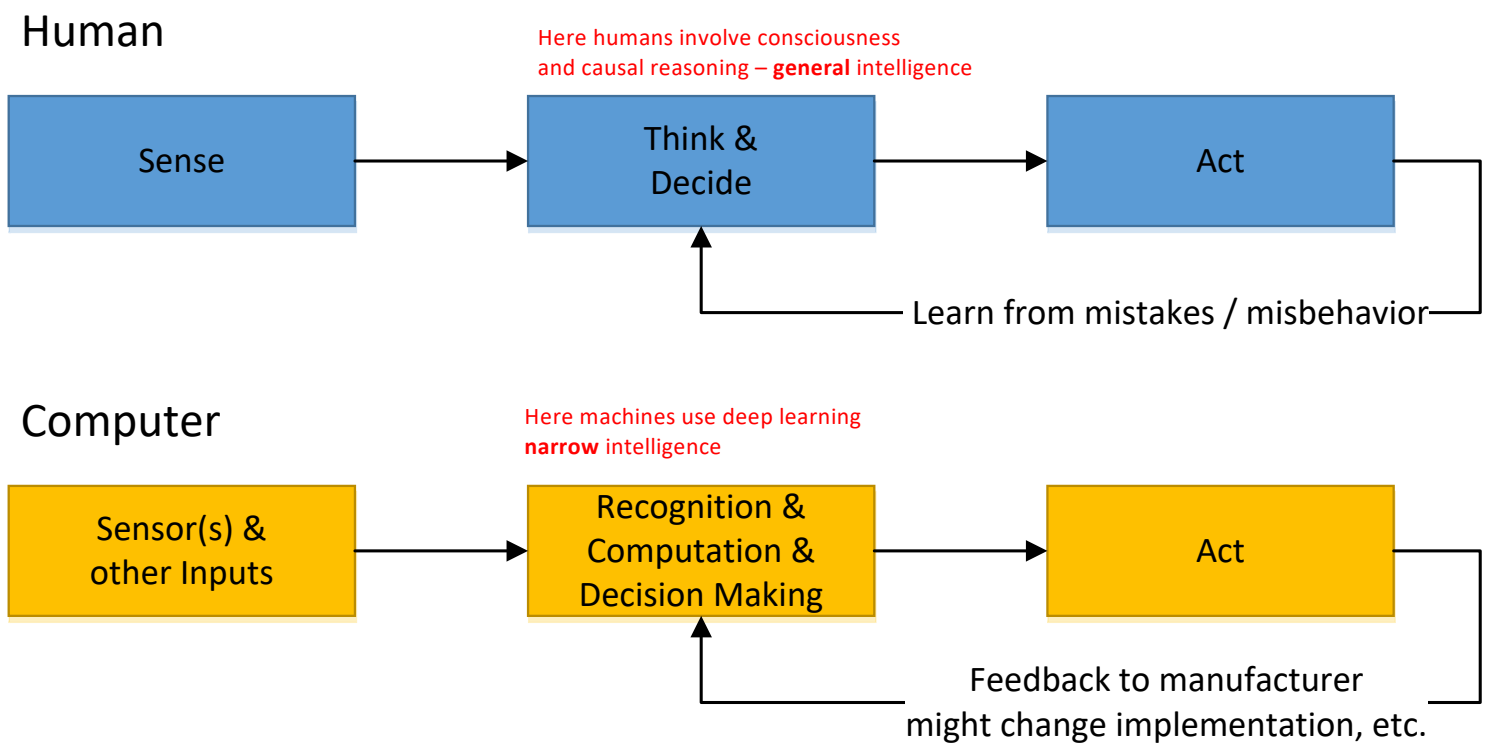


INTELLIGENCE OF AUTONOMOUS CARS – NARROW AI

(FAR FROM HUMAN LEVEL GENERAL AI)



Human Decision-making Process versus Self-Driving Car (Computer)



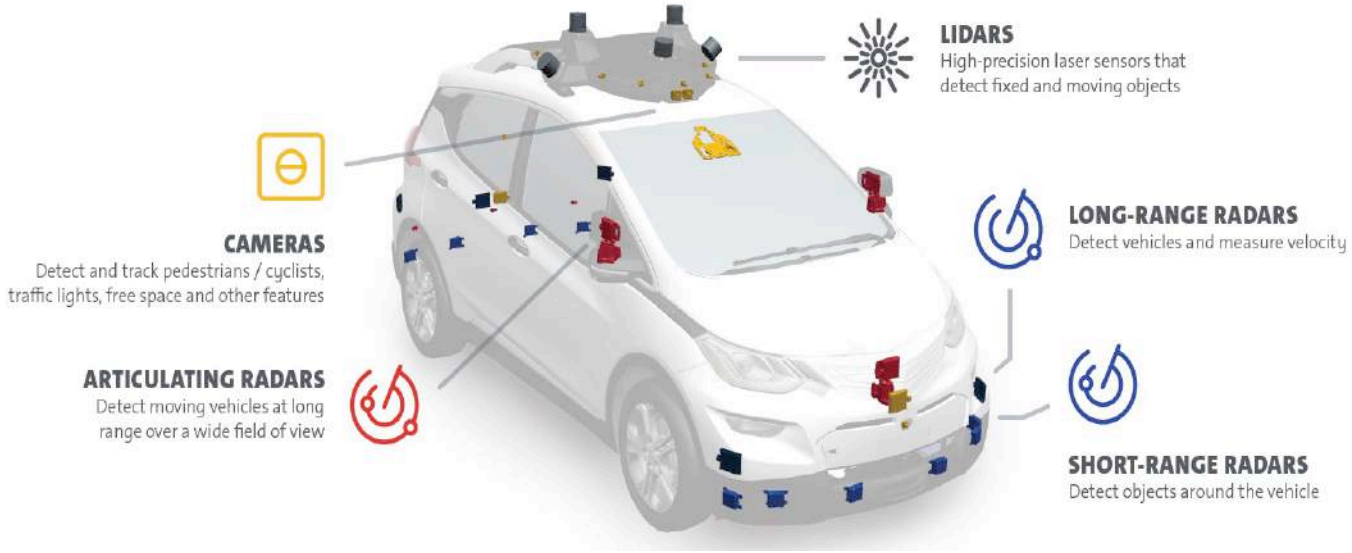
Decision Making in Self-Driving Cars

Decision making process involves sensors, external sources of information, networks, hardware, software, etc.

Environmental influences, such as weather conditions (rain, bright sun, storm, …)

Complex input must be filtered and only represents an abstraction of the real world.

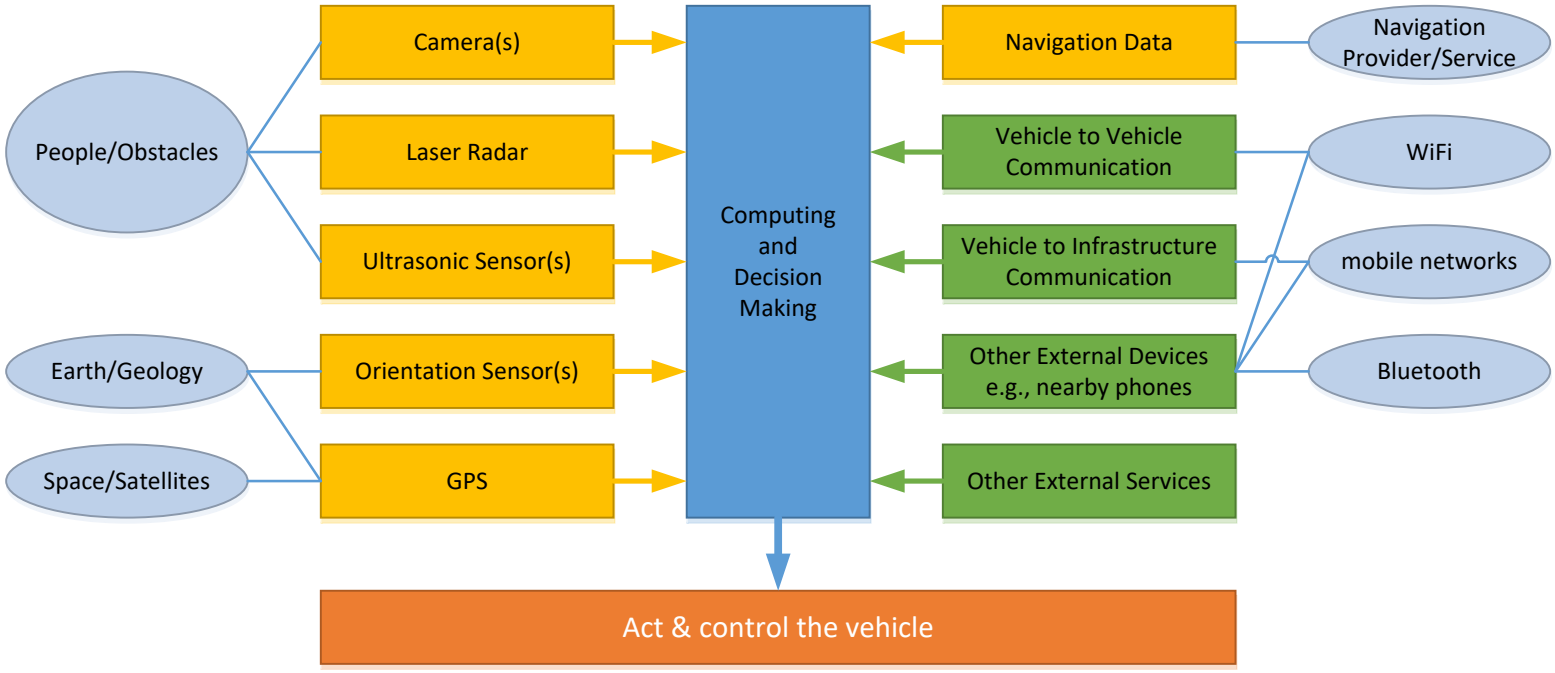
Technical Components



Picture Source: General Motors Safety Report 2018

What does a Self-Driving Car “see”…

Abstract Decision-Making Process



This is an outline of what a decision-making process might include.
Based on a literature review and official press releases (Tesla, Google, GM).

TECHNICAL CHALLENGES WITH ETHICAL CONSEQUENCES IN AUTONOMOUS CARS

Safety

How can we test self-driving cars?	and when is testing sufficient?
Real world vs Abstract world	Training of Neural Networks

Security

Attacks against car systems and sensors

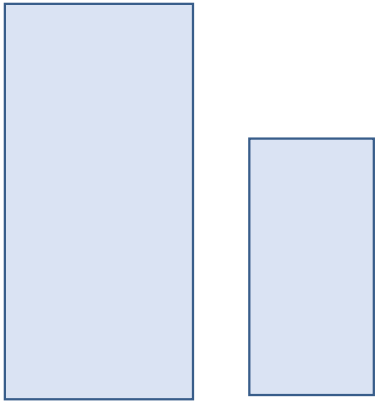
System & security updates

Do we need a “black box” in self-driving cars like in aircrafts?

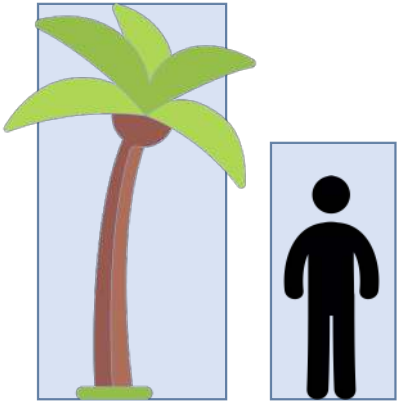
Privacy

- What data should the car have access to?
  - Who will have access to that data?
  - How will the data be used?
- What data is collected?

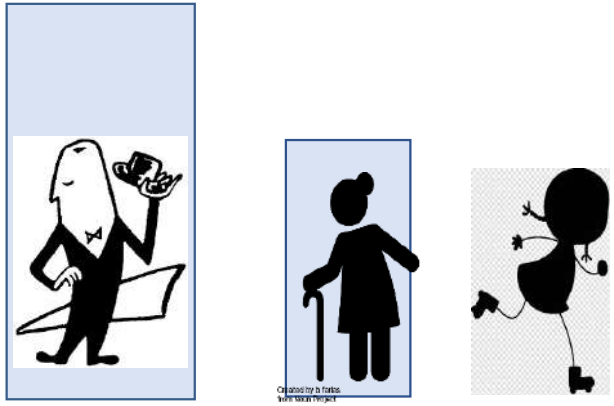
Privacy
What does the car “recognize” ?- Equivalent for Intelligent Robotics!



Objects, different size,
Position, moving or stationary



Objects vs Person(s)



„Everything“ including human identity – connected to data-bases

Trust

How trustworthy are data sources?

E.g., GPS, map data, external services

Trust between self-driving car and services

How trustworthy is the self-driving car?

E.g., Trust between user and car

Transparency

Multi-disciplinary challenge to ensure transparency, while respecting intellectual property rights, corporate secrets, security concerns, etc.

How much should be disclosed, and disclosed to whom?

Reliability

What do we have to rely on?

- What if sensor(s) fail?
- What if networks fail?

Redundancy for everything?

Responsibility and Accountability

- Who is responsible and for what?
- Who is accountable and for what?
- How is responsibility distributed among:
  - Developers
  - Car manufacturers
  - Safety inspectorates
  - Governmental institutions
  - Involved participants in the traffic
  - Other stakeholders

Quality Assurance Process

- Lifetime of components
- Maintenance
- Ethics-aware decision making in all processes will help to make ethically justified decisions.

SOCIAL CHALLENGES WITH ETHICAL CONSEQUENCES IN AUTONOMOUS CARS

Equivalent for Intelligent Social Robots will be Central!

Stakeholders Interests

- Loss of jobs (for cabs/taxi/truck/heavy industrial vehicles drivers)
- Humans in the loop
- Impact on Society

Stakeholders Interests

Freedom of movement

Will the car go, where I want it to go?

Implementation of restrictions

Route to Destination

Can the passenger define the route, or is it determined by the system?

Road trips?

Addressing Organisational Ethical Issues of AI

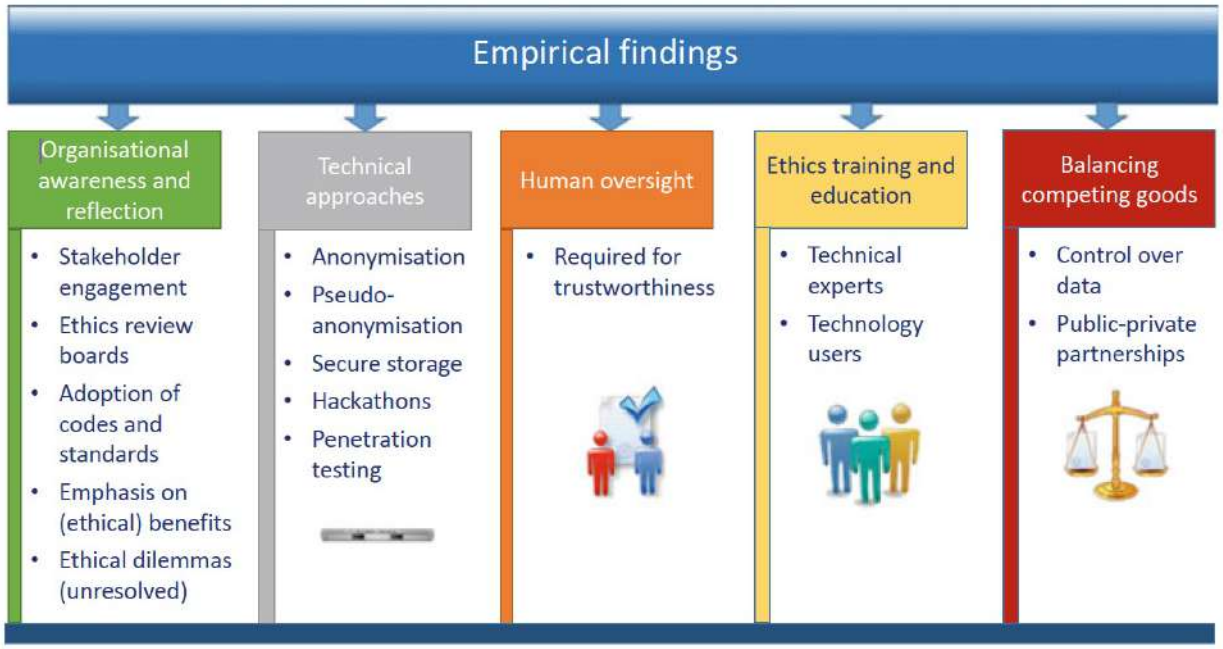
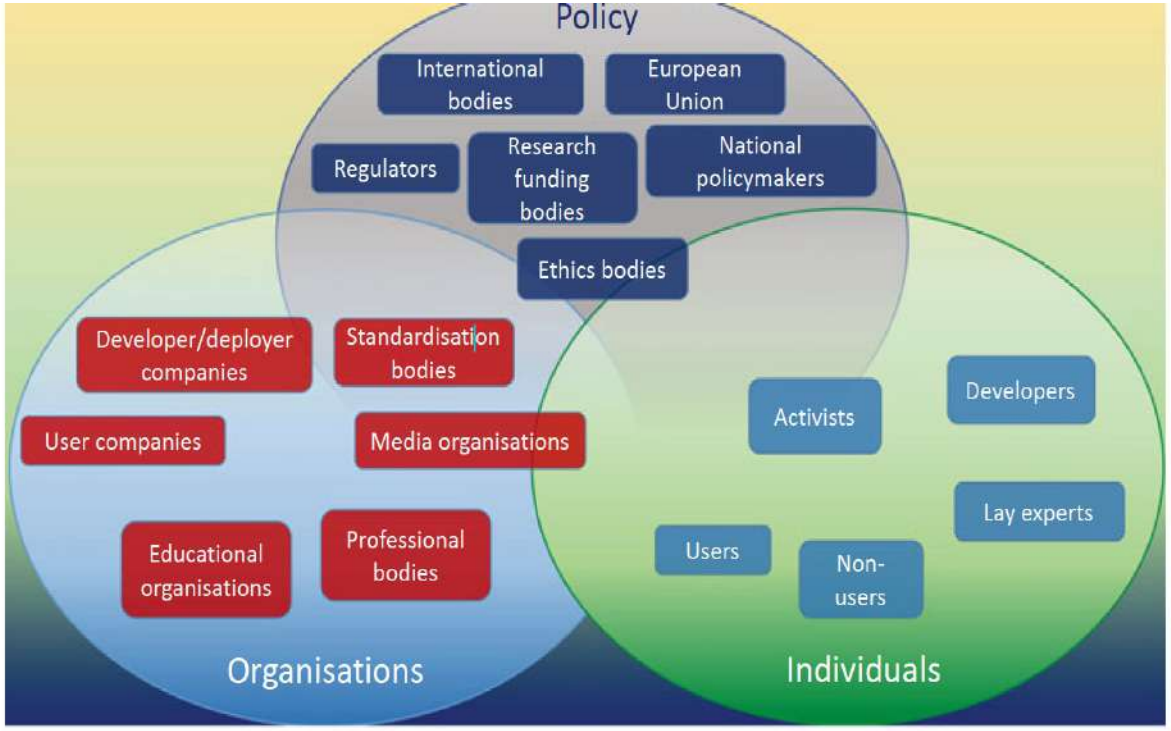


Fig. 5.2 How case study organisations address ethical issues of AI: empirical findings

Bernd Carsten Stahl (2021) Artificial Intelligence for a Better Future, https://link.springer.com/book/10.1007%2F978-3-030-69978-9

Overview of AI stakeholders, Artificial Intelligence for a Better Future



Bernd Carsten Stahl (2021) Artificial Intelligence for a Better Future, https://link.springer.com/book/10.1007%2F978-3-030-69978-9

KEY CHALLENGES OF ETHICAL GOVERNANCE OF AI SYSTEMS

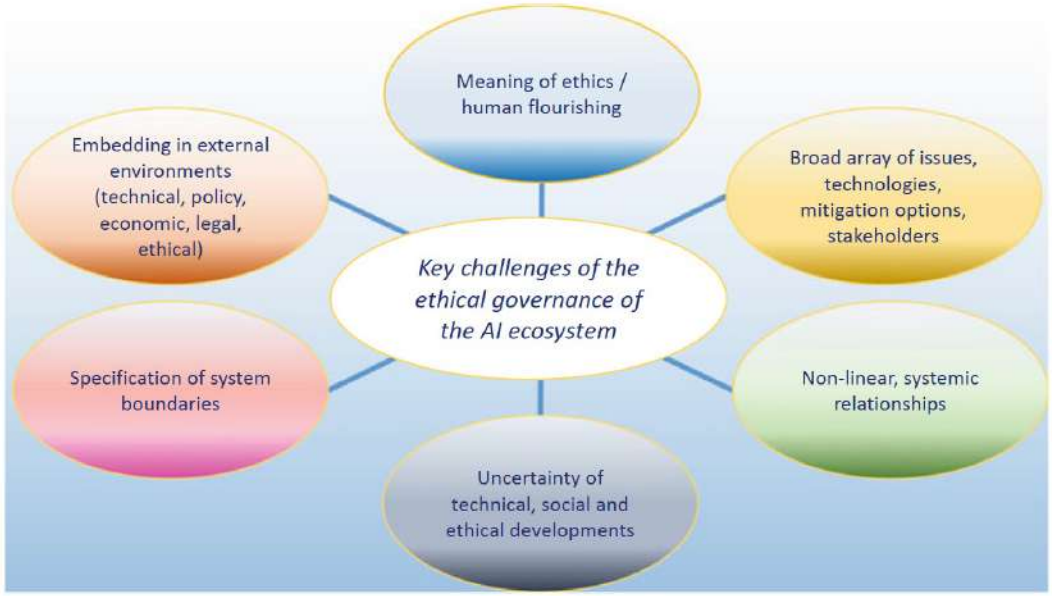


Fig. 7.1 Key challenges of ethical governance of AI ecosystems

Bernd Carsten Stahl (2021) Artificial Intelligence for a Better Future, https://link.springer.com/book/10.1007%2F978-3-030-69978-9

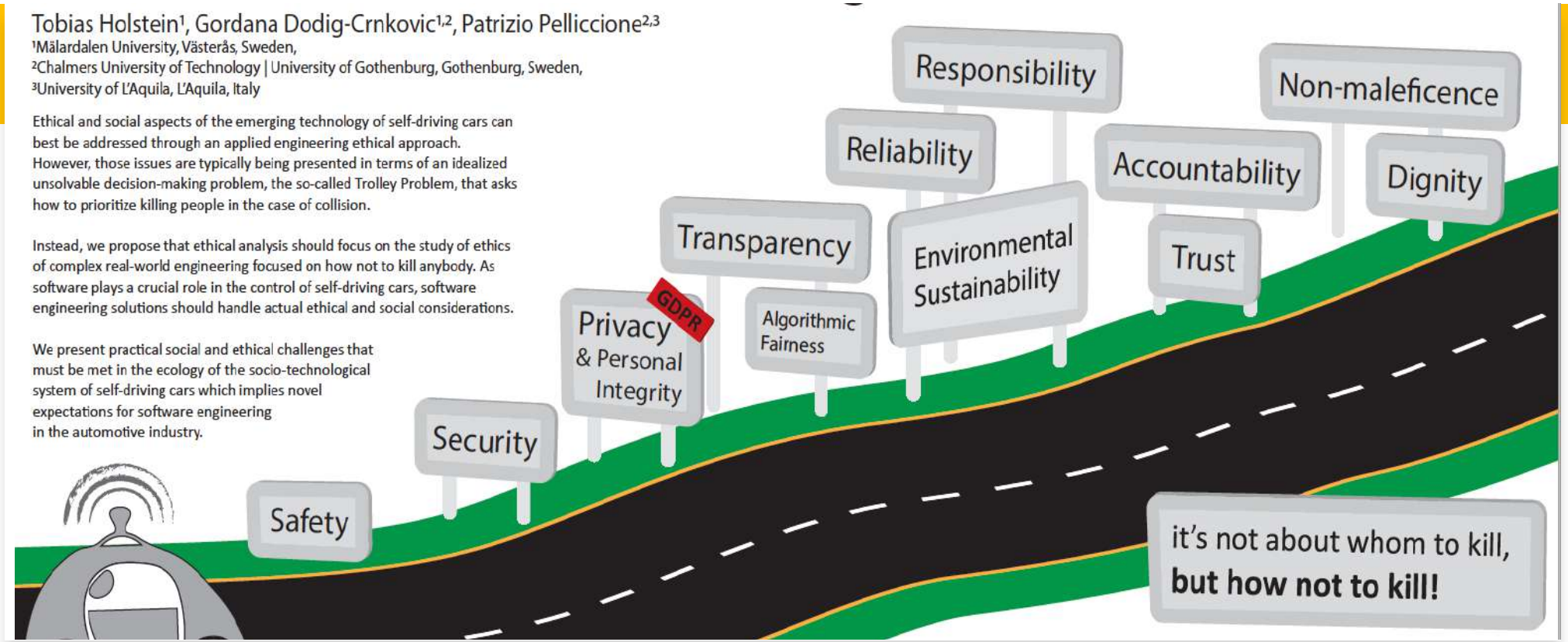
Ethical Issues of AI

Table 4.1 Three categories of ethical issues of artificial intelligence

1. Issues arising from machine learning	
Privacy and data protection	Lack of privacy
	Misuse of personal data
	Security problems
Reliability	Lack of quality data
	Lack of accuracy of data
	Problems of integrity
Transparency	Lack of accountability and liability
	Lack of transparency
	Bias and discrimination
	Lack of accuracy of predictive recommendations
	Lack of accuracy of non-individual recommendations
Safety	Harm to physical integrity
2. Living in a digital world	
Economic issues	Disappearance of jobs
	Concentration of economic power
	Cost to innovation
Justice and fairness	Contested ownership of data
	Negative impact on justice system
	Lack of access to public services
	Violation of fundamental human rights of end users
	Violation of fundamental human rights in supply chain
	Negative impact on vulnerable groups
	Unfairness
Freedom	Lack of access to and freedom of information
	Loss of human decision-making
	Loss of freedom and individual autonomy
Broader societal issues	Unequal power relations
	Power asymmetries
	Negative impact on democracy
	Problems of control and use of data and systems
	Lack of informed consent
	Lack of trust
	Potential for military use
	Negative impact on health
	Reduction of human contact
	Negative impact on environment
Uncertainty issues	Unintended, unforeseeable adverse impacts
	Prioritisation of the "wrong" problems
	Potential for criminal and malicious use
3. Metaphysical issues	
	Machine consciousness
	"Awakening" of AI
	Autonomous moral agents
	Super-intelligence
	Singularity
	Changes to human nature

https://link.springer.com/book/10.1007%2F978-3-030-69978-9 Artificial Intelligence for a Better Future (book)

Ethical Guidelines for Self-Driving Cars



Ehics Of Self-Driving Cars

Presented at major SE conference ICSE2020 as poster

Extended version in a book chapter:

Holstein, T., Dodig-Crnkovic, G., & Pelliccione, P. (2021). Steps Towards Real-world Ethics for Self-driving Cars: Beyond the Trolley Problem. In Steven John Thompson (Ed.), Machine Law, Ethics, and Morality in the Age of Artificial Intelligence. IGI Global

Conclusions

It is time to stop discussing unsolvable ethical dilemmas that obfuscate much bigger actual ethical challenges of technology.

Discuss the real-world ethical challenges surrounding emerging technology.

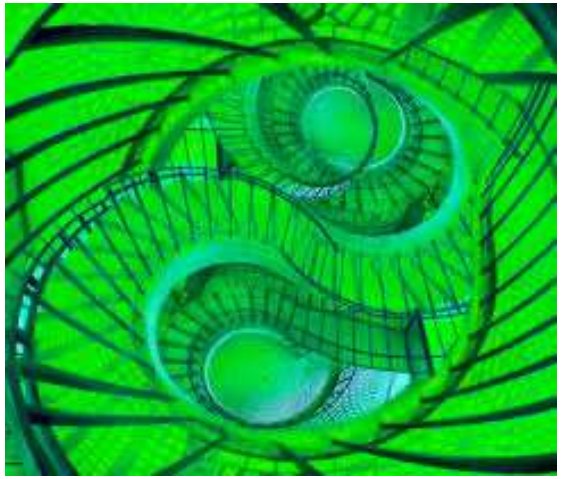
Define what is technically possible and ethically justifiable.

Create transparency to support evaluations by independent organizations/experts.

Ethicality/Ethicity as non-functional property? (Ethicality: the state, quality, or manner of being ethical.)

There is already a body of normative documents that can support ethicality of design and implementation.

Avslutning och diskussion



Tankar om AI risker och om vikten att reglera AI:
https://www.youtube.com/watch?v=QEGjCcU0FLs
Will AI Be Humanity's Last Act? with Stuart Russell
(författaren till "AI boken")

https://www.apogeonline.com/articoli/regole-di-composizione-fotografica-usare-la-prospettiva-a-piu-punti-harold-davis/

FURTHER READING

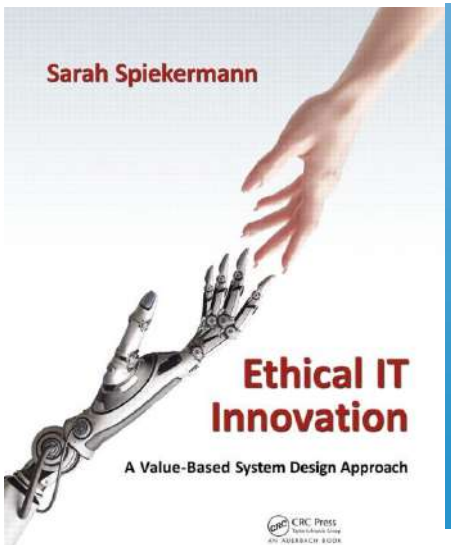
81

Challenges

Legislation	Global framework	Guidelines	Including Ethics into all phases in the life-cycle
Keeping legislation up-to-date with current level of automated driving, and emergence of self-driving cars	Creating and defining global legislation frameworks for the implementation of interoperable and development of increasingly automated vehicles	Defining the guidelines that will be adopted by society for building self-driving cars	Including ethical guidelines in design, development and other processes in the life-cycle of a product.

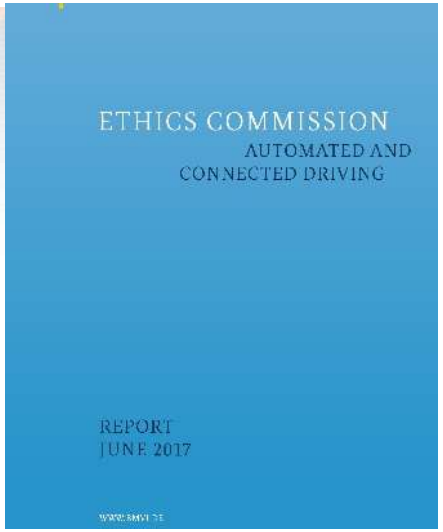
Holstein, Dodig-Crnkovic, Pellizzione: Ethical and Social Aspects of Self-Driving Cars, ArXives

A Value-Based Design Approach



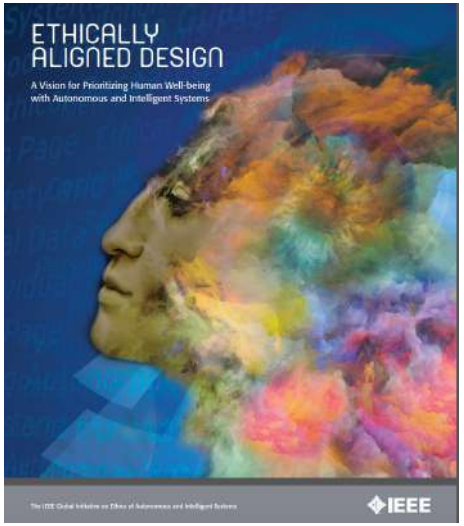
Sarah Spiekermann

Ethical IT Innovation:
A Value-Based System Design Approach



Ethics Commission: Automated and connected driving (Report by Federal Ministry of Transport and Digital Infrastructure of Germany [BMVI])

BMVI = Bundesministerium für Verkehr und digitale Infrastruktur



https://ethicsinaction.ieee.org/