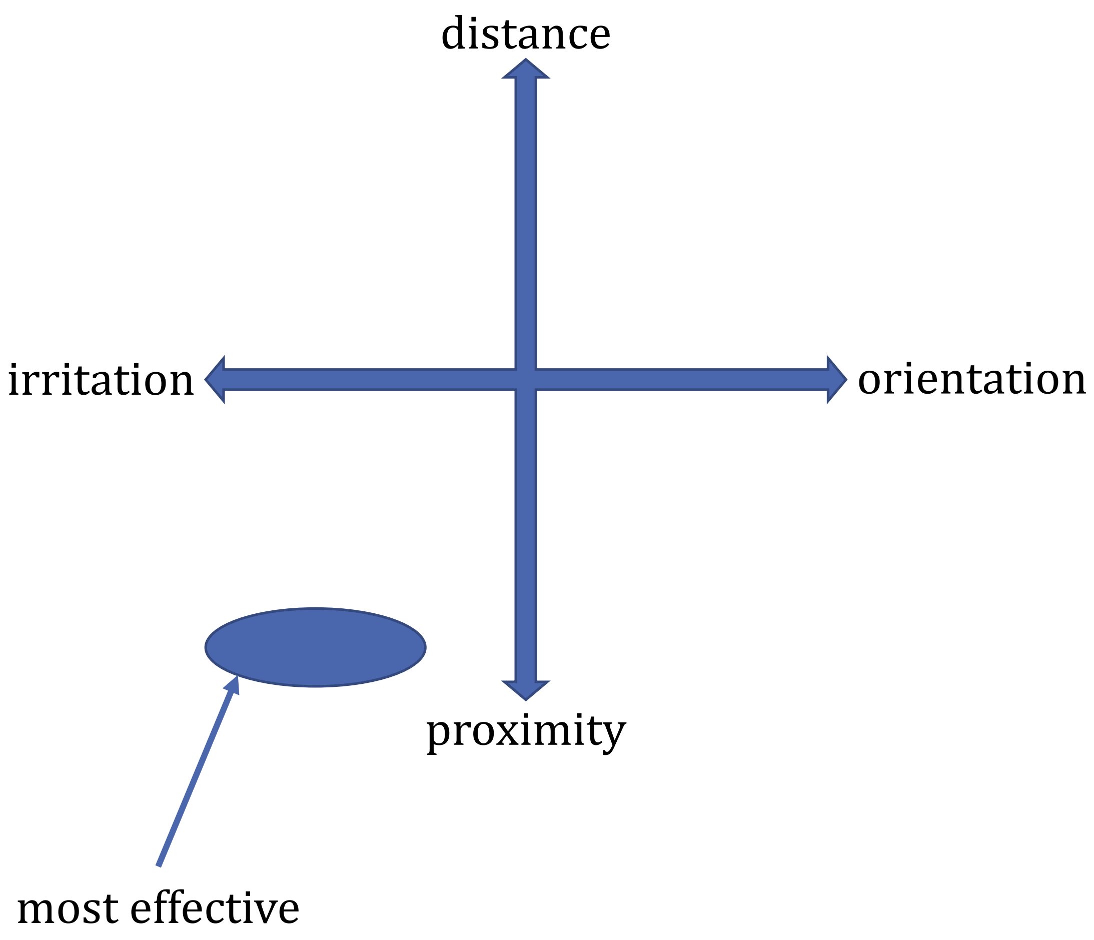
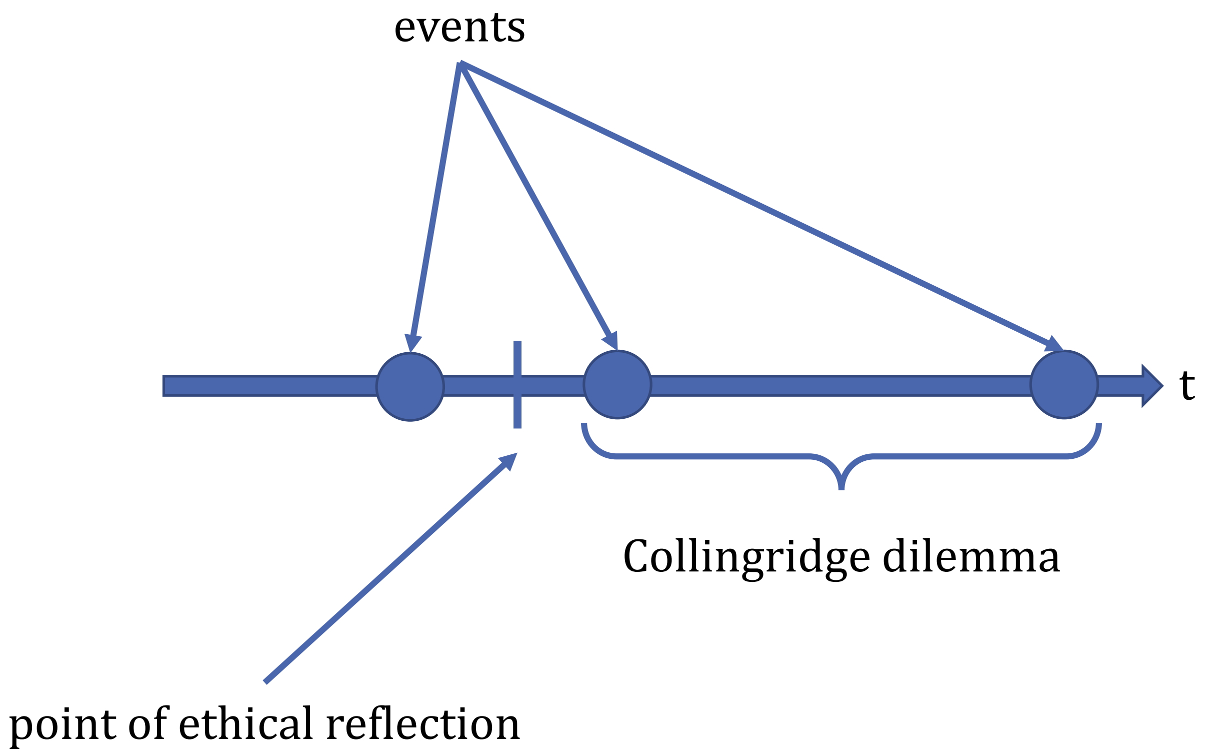
Reflecting opportunities and risks – AI ethics and its broad range of issues

Dr. Thilo Hagendorff

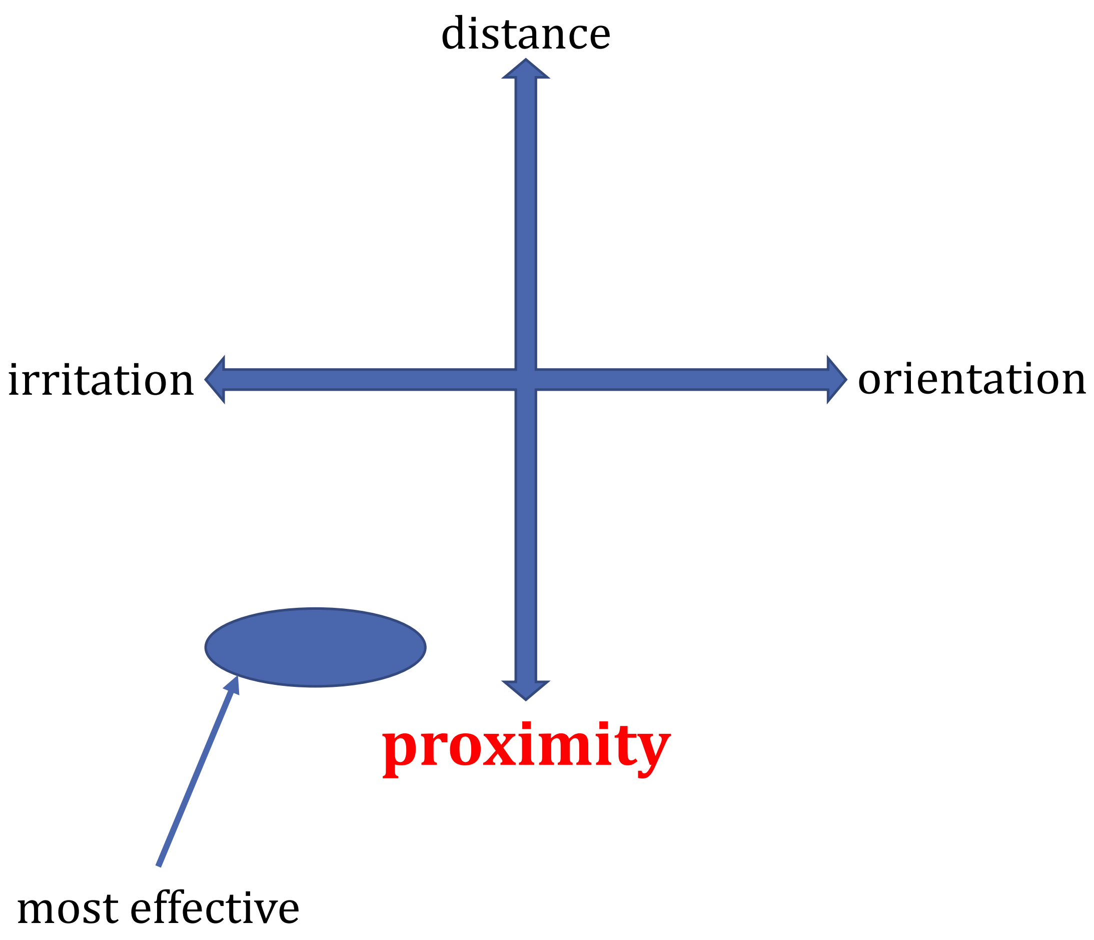
University of Tuebingen
Cluster of Excellence Machine Learning

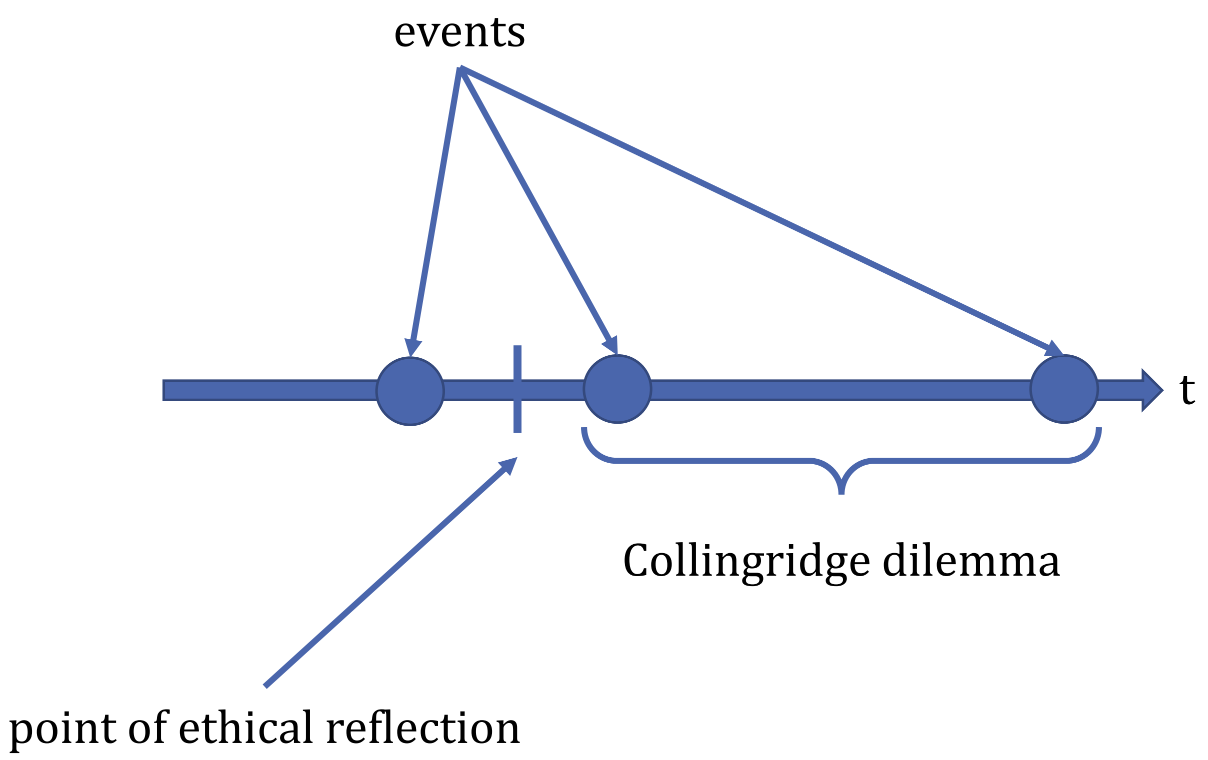
Dimensions of AI ethics



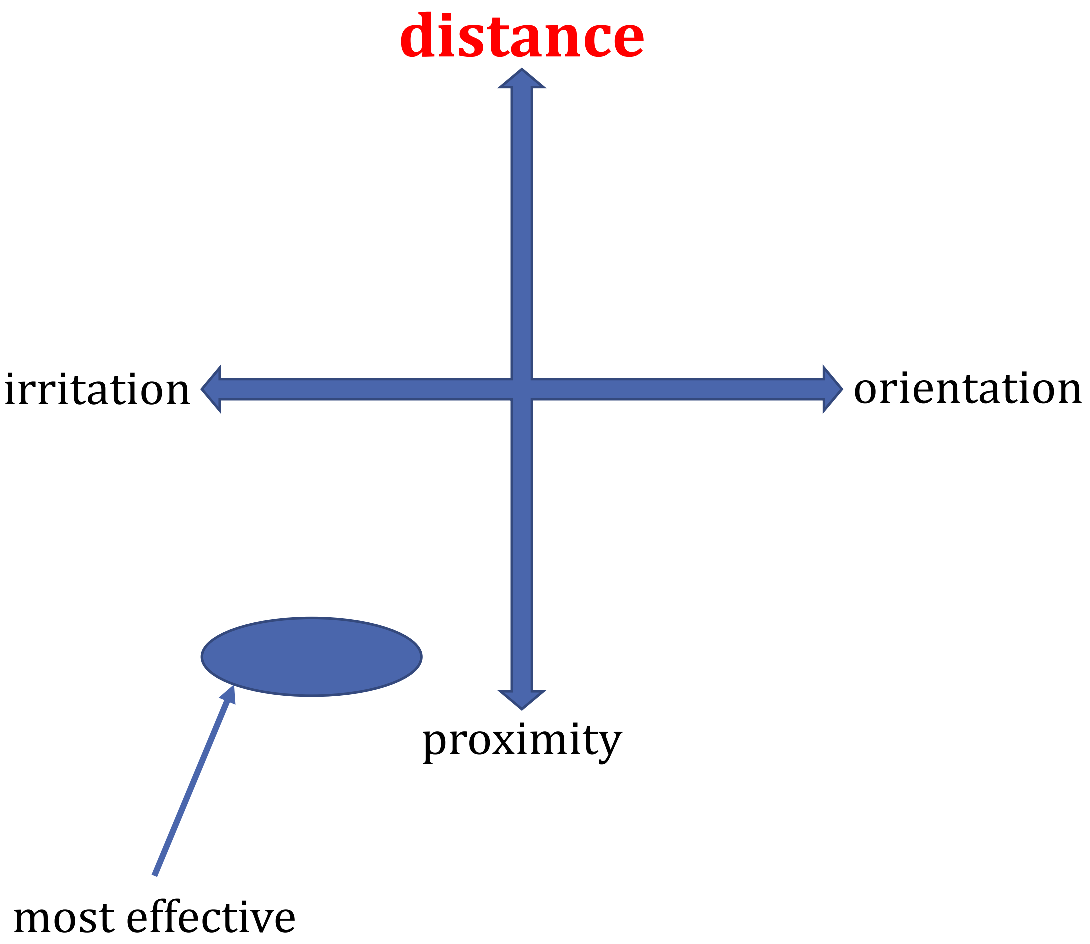


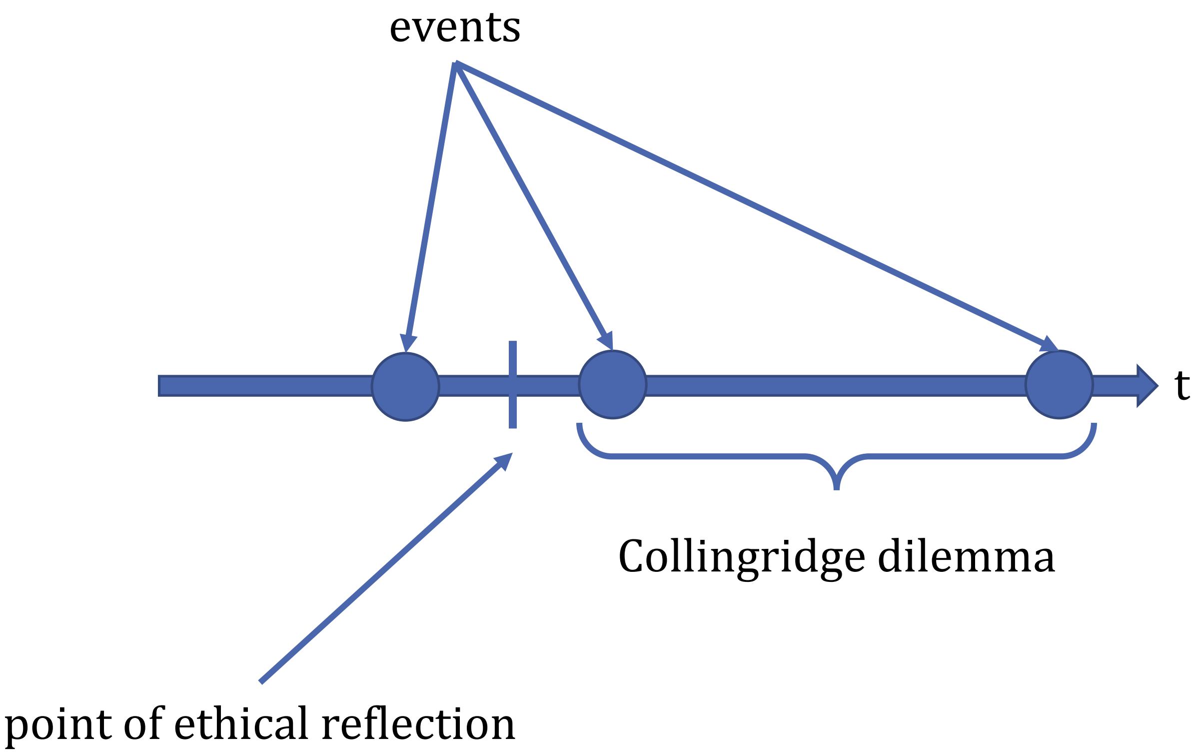
Dimensions of AI ethics



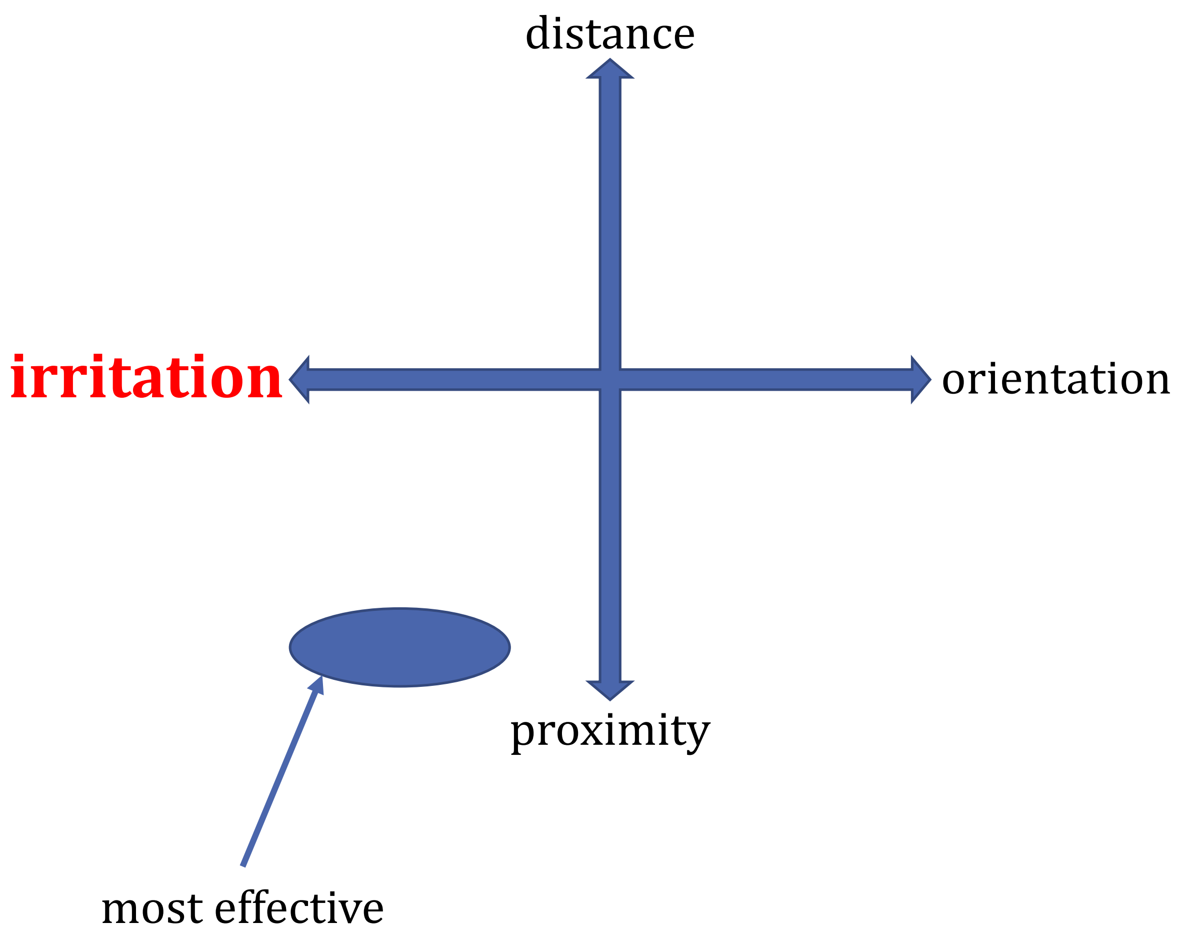


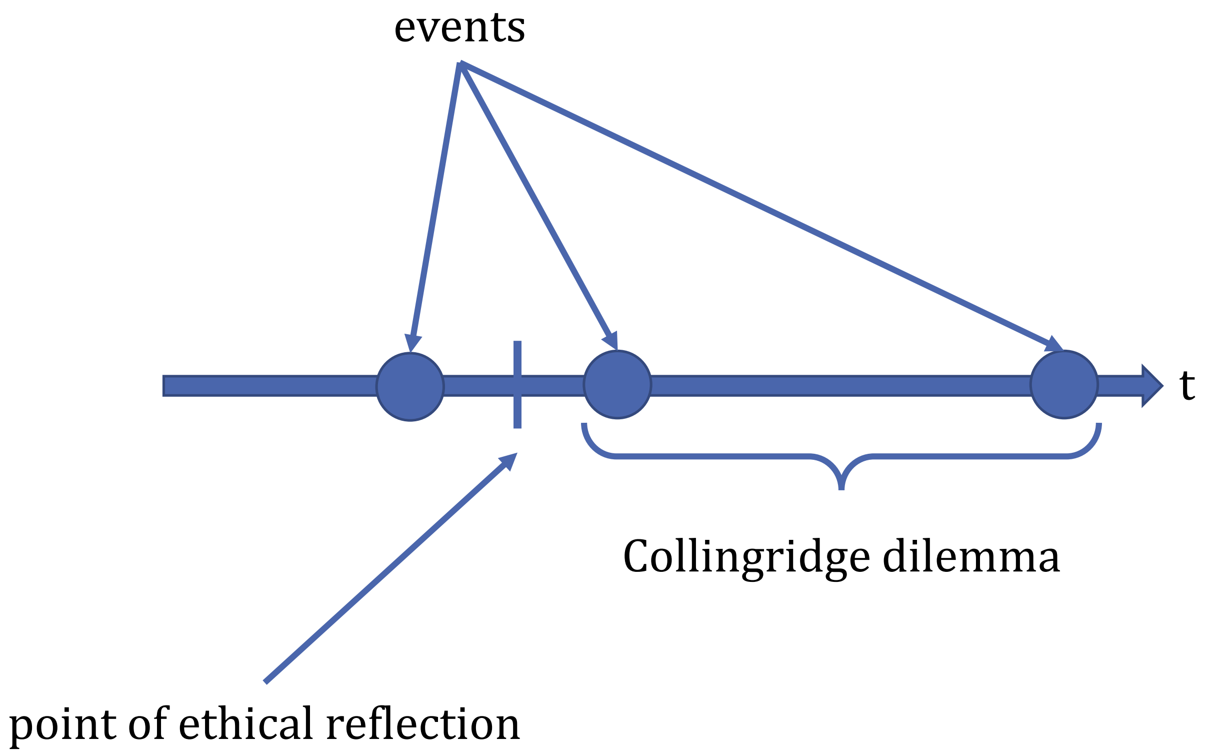
Dimensions of AI ethics



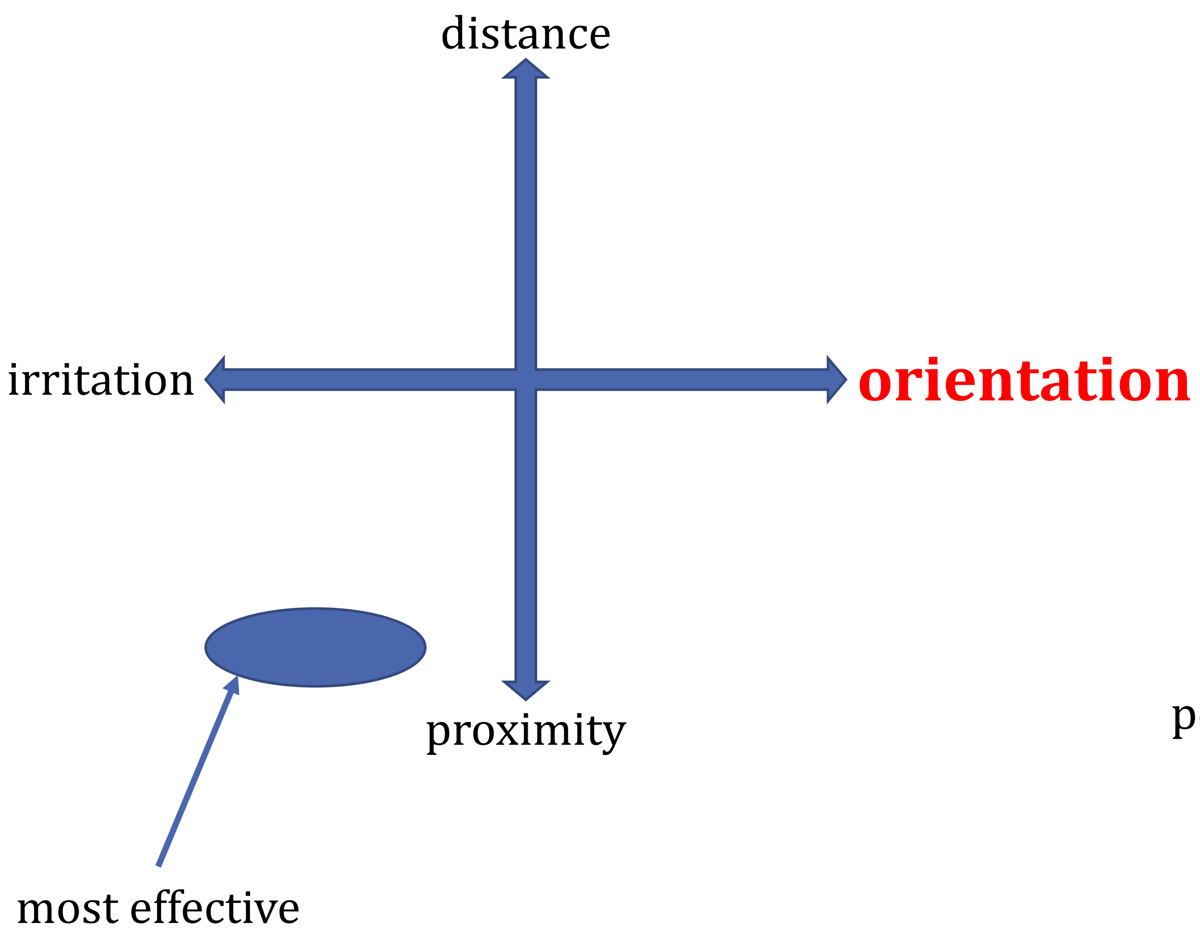


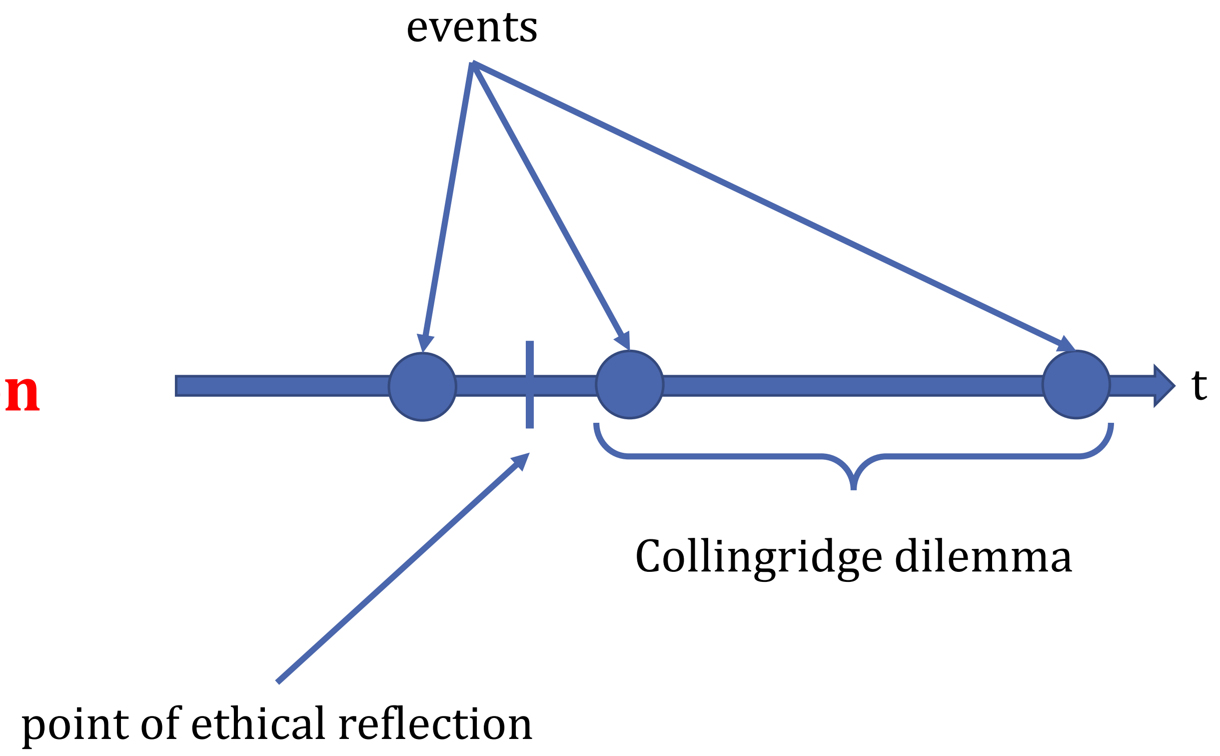
Dimensions of AI ethics



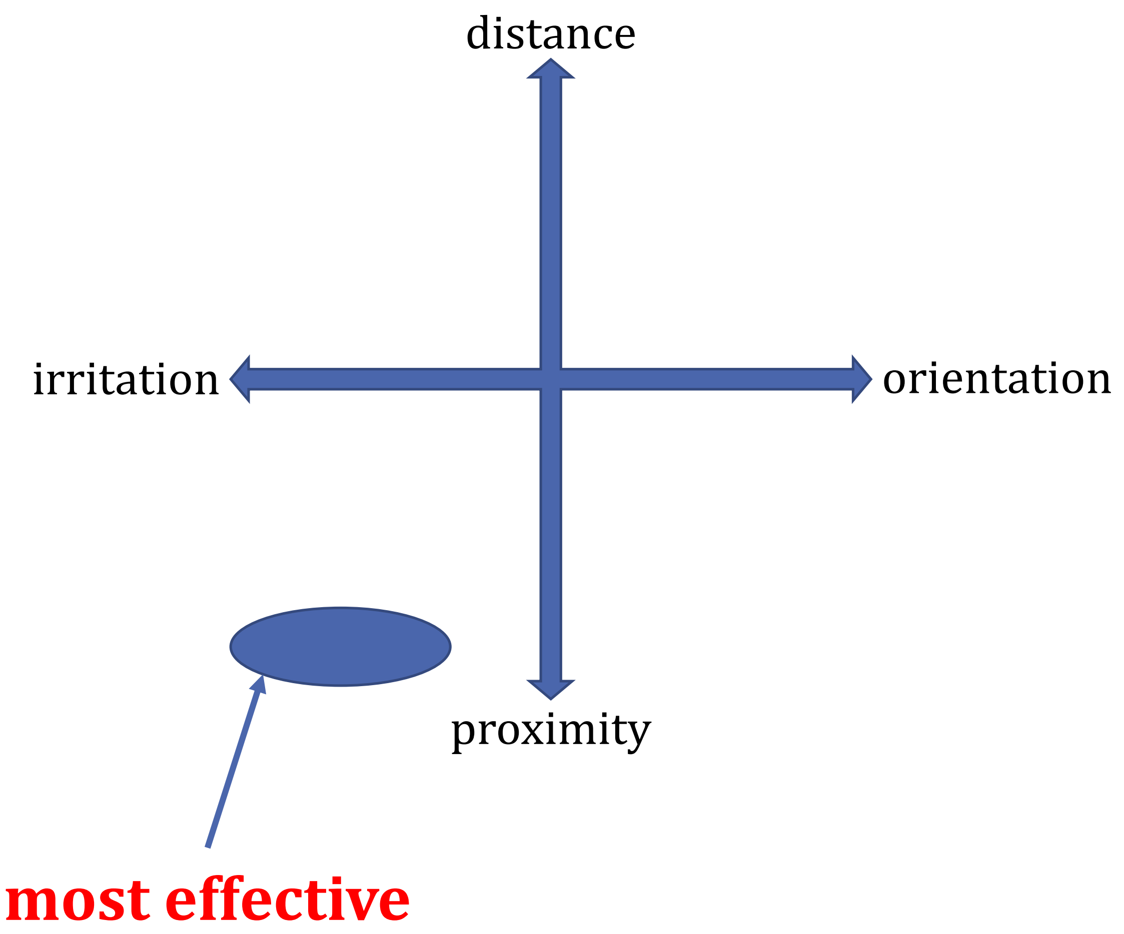


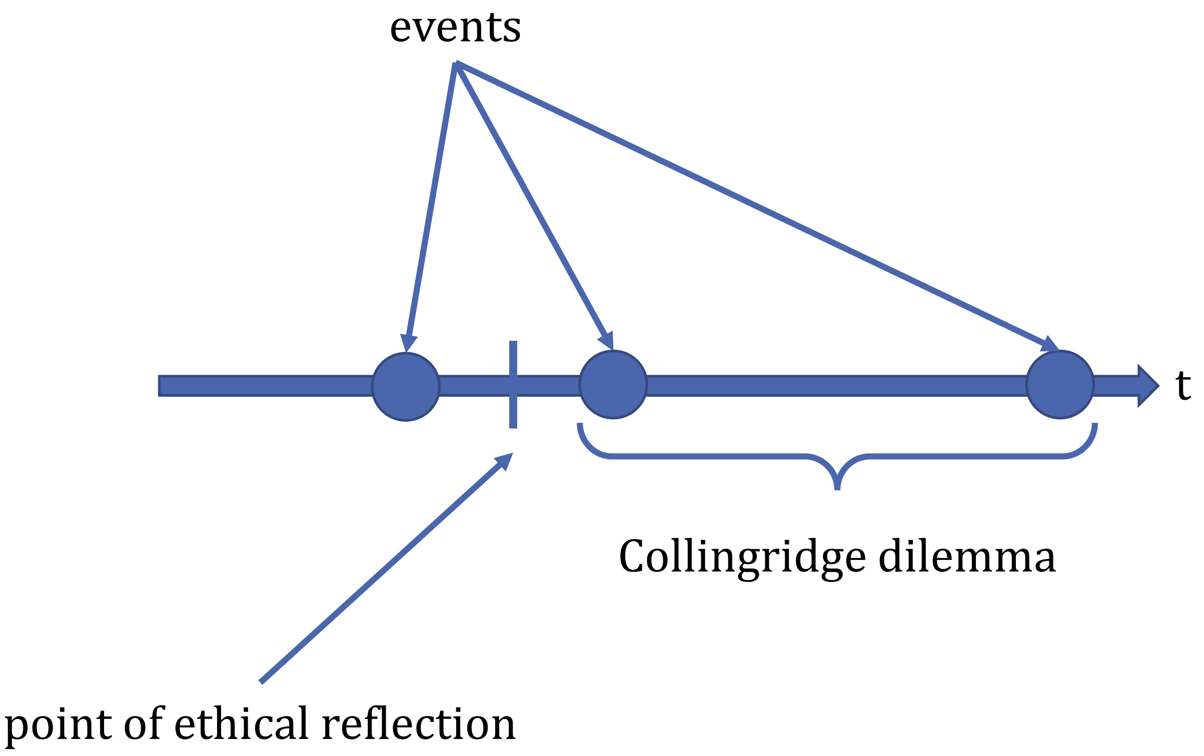
Dimensions of AI ethics



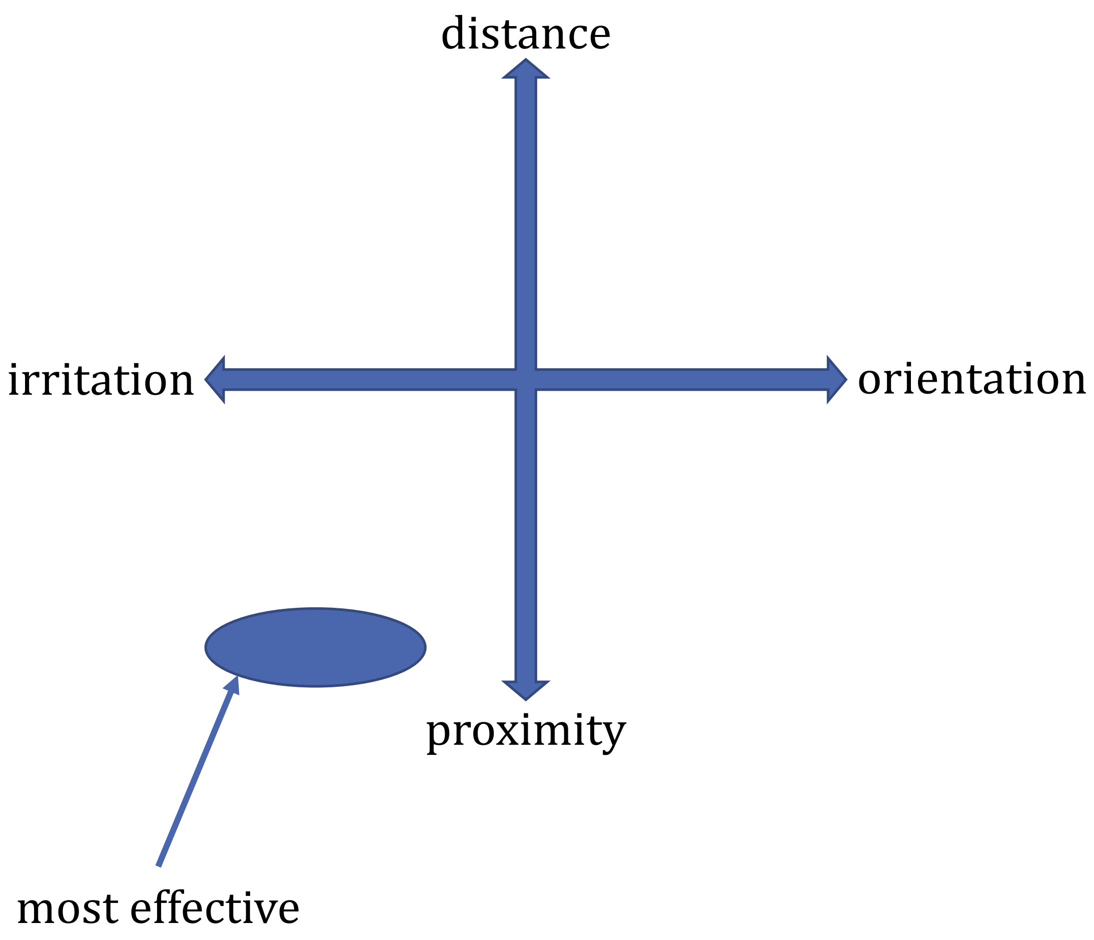


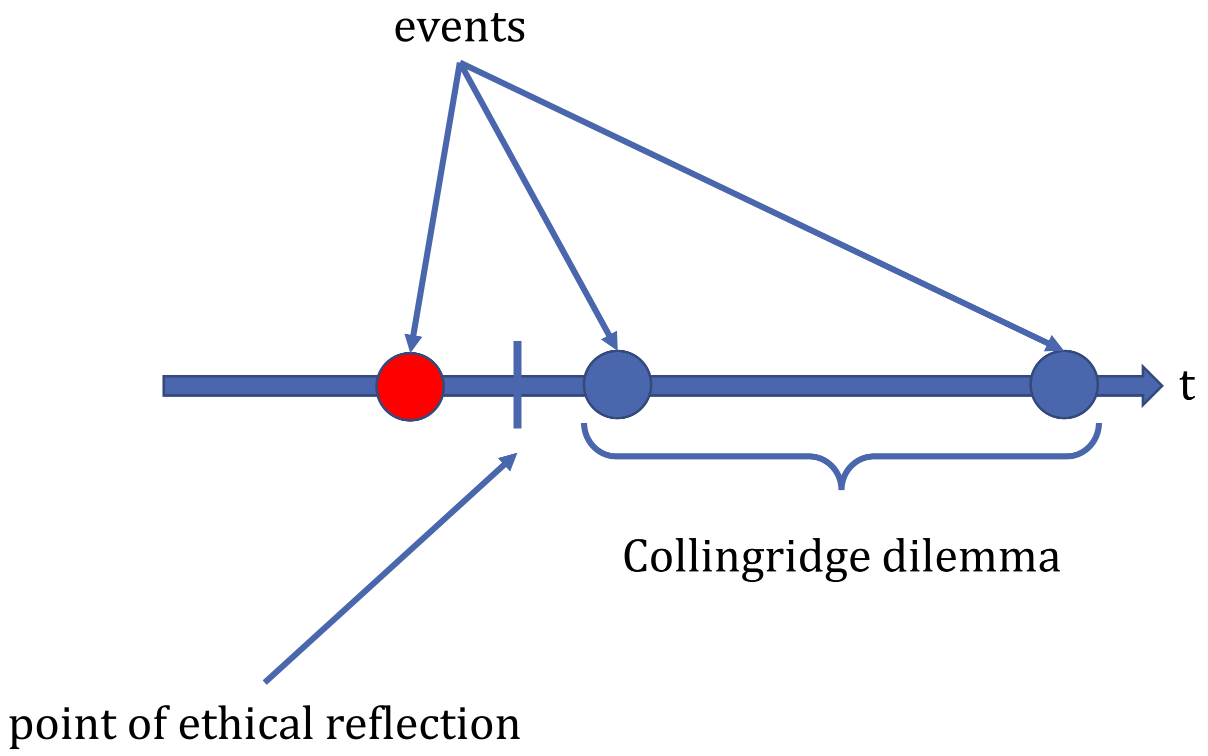
Dimensions of AI ethics



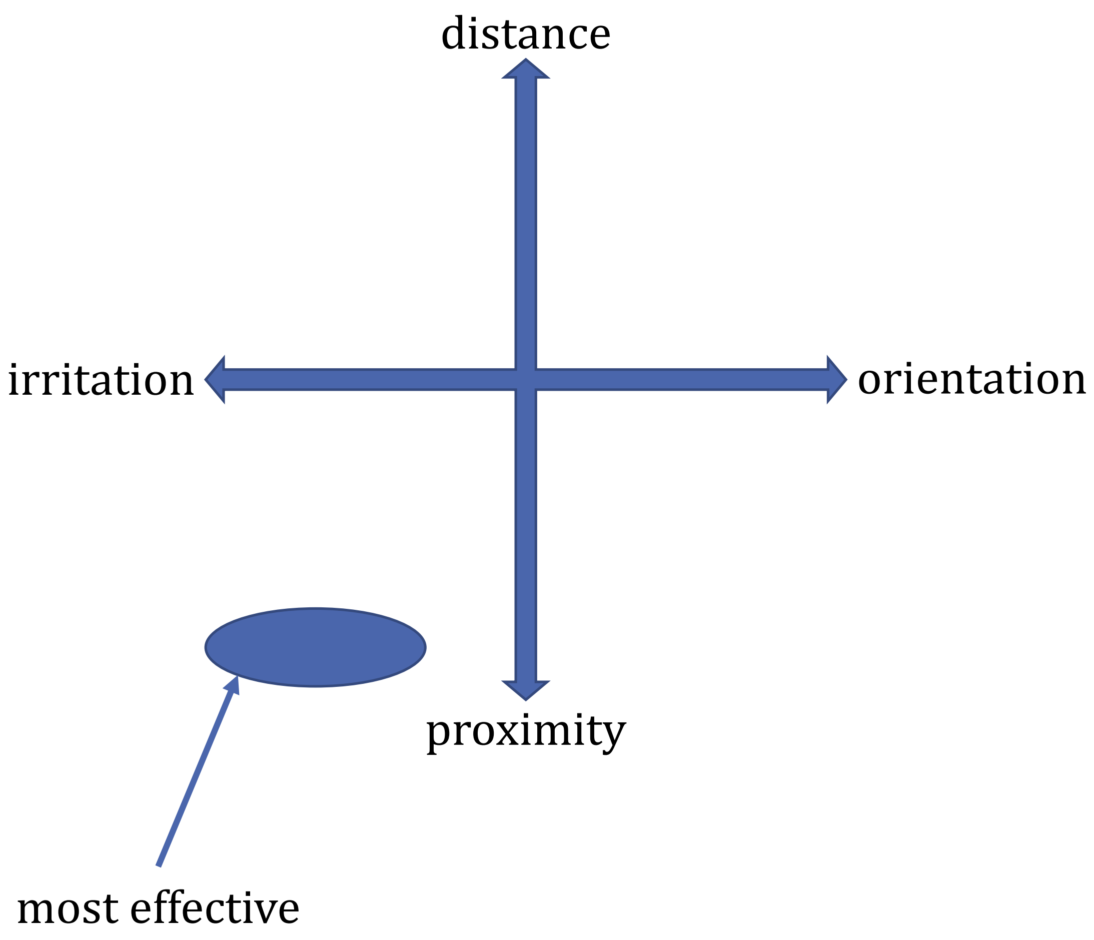


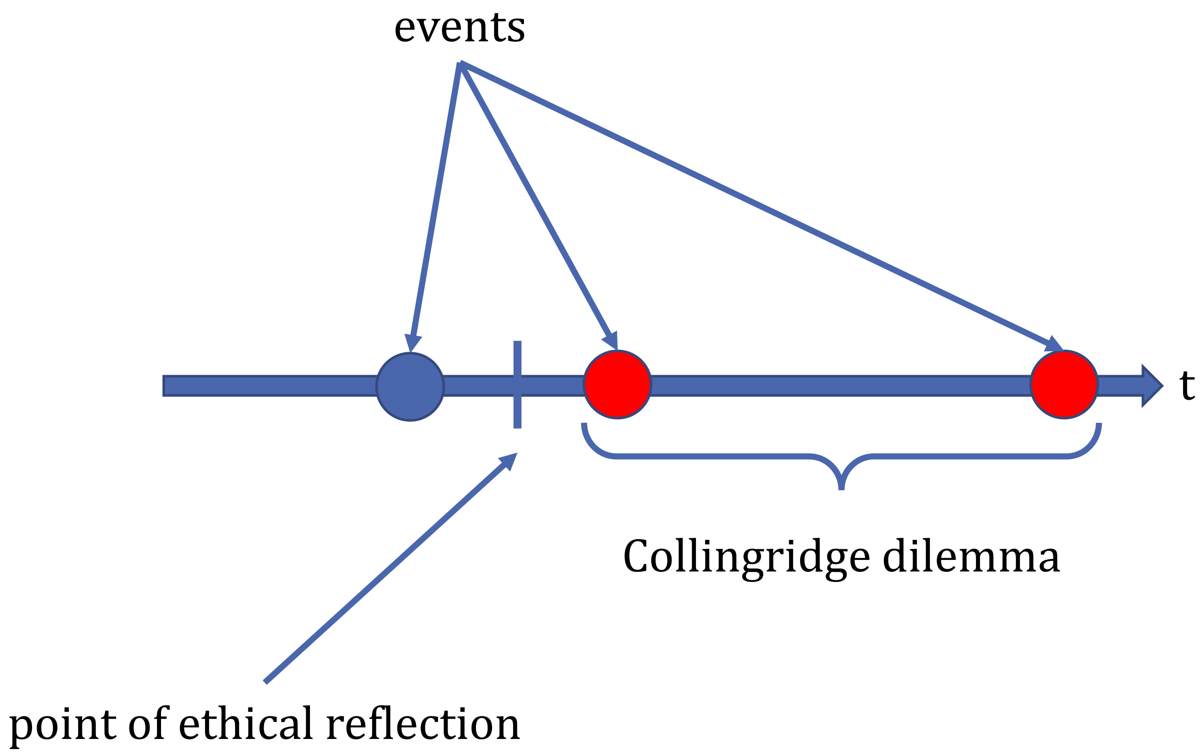
Dimensions of AI ethics





Dimensions of AI ethics





Demands for AI/ML Ethics

TC

The future of AI relies on a code of ethics

Matthew Howard @mattdhoward / 10 months ago

WIRED SUBSCRIBE

TOM SIMONITE BUSINESS 05.16.18 04:32 PM

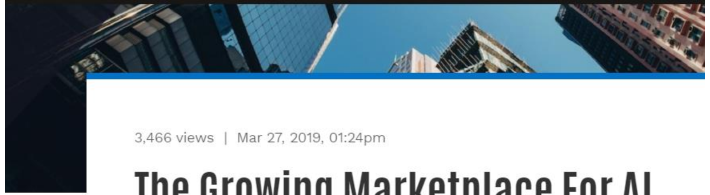
TECH FIRMS MOVE TO PUT ETHICAL GUARD RAILS AROUND AI

NETZPOLITIK.ORG

Technologie

Keine roten Linien: Industrie entschärft Ethik-Leitlinien für Künstliche Intelligenz

Forbes

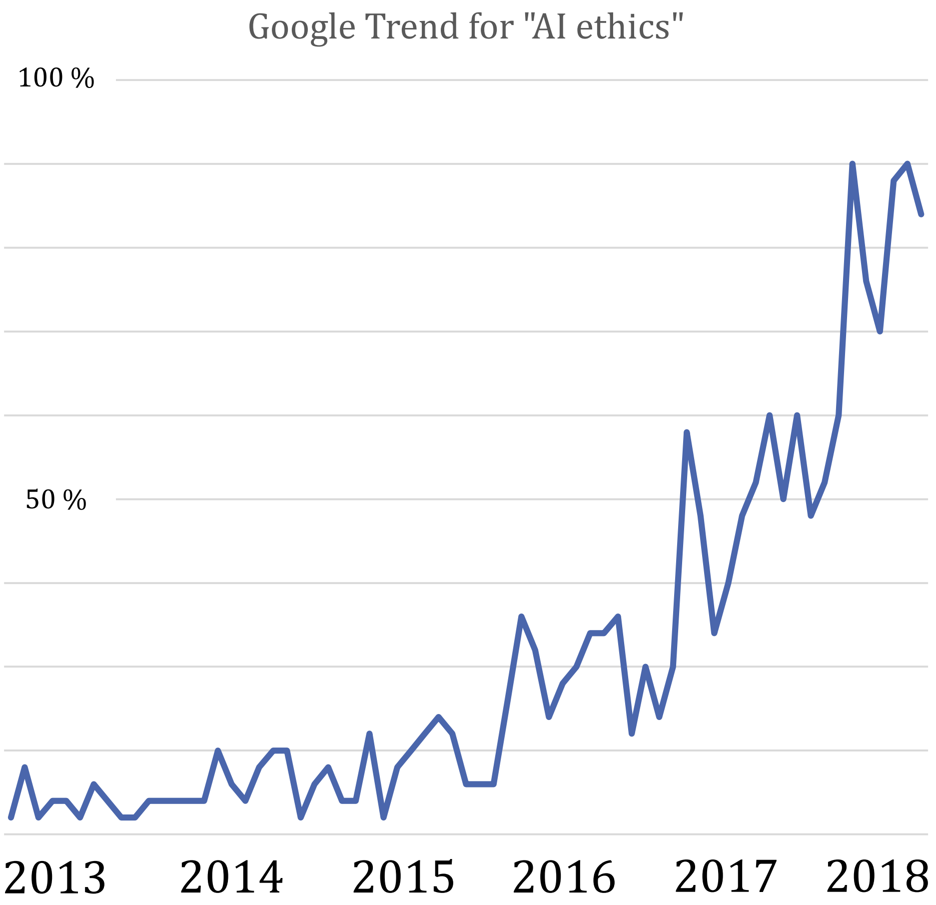


3,466 views | Mar 27, 2019, 01:24pm

The Growing Marketplace For AI Ethics

Forbes Insights

Insights Team Insights Contributor
FORBES INSIGHTS With Intel AI



Guidelines

- Hagendorff, Thilo (2019): The Ethics of AI Ethics. An Evaluation of Guidelines. in: arXiv:1903.03425v1, pp. 1–15.

The Ethics of AI Ethics

An Evaluation of Guidelines

Dr. Thilo Hagendorff

University of Tuebingen
International Center for Ethics in the Sciences and Humanities
thilo.hagendorff@uni-tuebingen.de

Abstract - Current advances in research, development and application of artificial intelligence (AI) systems have yielded a far-reaching discourse on AI ethics. In consequence, a number of ethics guidelines have been released in recent years. These guidelines comprise normative principles and recommendations aimed to harness the "disruptive" potentials of new AI technologies. Designed as a comprehensive evaluation, this paper analyzes and compares these guidelines highlighting overlaps but also omissions. As a result, I give a detailed overview of the field of AI ethics. Finally, I also examine to what extent the respective ethical principles and values are implemented in the practice of research, development and application of AI systems – and how the effectiveness in the demands of AI ethics can be improved.

Keywords - artificial intelligence, machine learning, ethics, guidelines, implementation

1 Introduction

The current AI boom is accompanied by constant calls for applied ethics, which are meant to harness the "disruptive" potentials of new AI technologies. As a result, a whole body of ethical guidelines has been developed in recent years collecting principles, which technology developers should adhere to as far as possible. However, the critical question arises: Do those ethical guidelines have an actual impact on human decision-making in the field of AI and machine learning? The short answer is: No, most often not. This paper analyzes fifteen of the major AI ethics guidelines and issues recommendations on how to overcome the relative ineffectiveness of these guidelines.

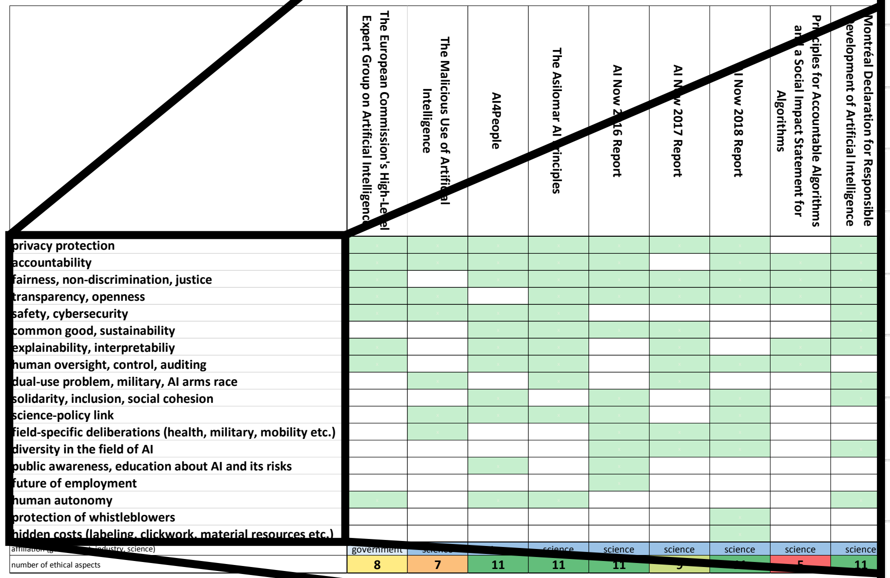
AI ethics – or ethics in general – lacks mechanisms to reinforce its own normative claims. Of course, the enforcement of ethical principles may involve reputational losses in the case of misconduct, or restrictions on memberships in certain professional bodies. Yet altogether, these mechanisms are rather weak and pose no eminent threat. Researchers, politicians, consultants, managers and activists have to deal with this essential weakness of ethics. However, it is also a reason why ethics is so appealing to many AI companies and institutions. When companies or research institutes formulate their own ethical guidelines, regularly incorporate ethical considerations into their public relations work, or adopt ethically motivated "self-commitments", efforts to create a truly binding legal framework are continuously discouraged. Ethics guidelines of the AI industry serve to suggest to legislators that internal self-governance in science and industry is sufficient, and that no specific laws are necessary to mitigate possible technological risks and to eliminate scenarios of abuse (Calo 2017). And even when more concrete laws concerning AI systems are demanded, as recently done by Google (Google 2019), these demands remain relatively vague and superficial.

Science- or industry-led ethics guidelines, as well as other concepts of self-governance, may serve to pretend that accountability can be devolved from state authorities and democratic institutions upon the respective sectors of science or industry. Moreover, ethics can also simply serve the purpose of calming critical voices from the public, while simultaneously the criticized practices are maintained within the organization. The association "Partnership on AI" (2018) which brings together companies such as Amazon, Apple, Baidu, Facebook, Google, IBM and Intel is exemplary in this context. Companies can highlight their membership in such associations whenever the

Guidelines

	The European Commission's High-Level Expert Group on Artificial Intelligence	The Malicious Use of Artificial Intelligence	AI4People	The Asilomar AI Principles	AI Now 2016 Report	AI Now 2017 Report	AI Now 2018 Report	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Montréal Declaration for Responsible Development of Artificial Intelligence	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems	ITI AI Policy Principles	Microsoft AI principles	Artificial Intelligence at Google	Everyday Ethics for Artificial Intelligence	Partnership on AI	number of mentions
privacy protection	x	x	x	x	x	x	x		x	x	x	x	x	x	x	**14**
accountability	x	x	x	x	x		x	x	x	x	x	x		x	x	**13**
fairness, non-discrimination, justice	x		x	x	x	x	x	x	x	x		x	x	x	x	**13**
transparency, openness	x	x		x	x	x	x	x	x	x		x				**10**
safety, cybersecurity	x	x	x	x					x	x	x	x	x		x	**10**
common good, sustainability			x	x	x	x			x	x			x	x	x	**9**
explainability, interpretabiliy	x		x	x		x		x	x		x			x		**8**
human oversight, control, auditing	x		x	x		x	x	x			x		x			**8**
dual-use problem, military, AI arms race		x		x		x			x	x			x			**6**
solidarity, inclusion, social cohesion			x		x		x		x			x			x	**6**
science-policy link		x	x	x	x		x									**5**
field-specific deliberations (health, military, mobility etc.)		x			x	x	x			x						**5**
diversity in the field of AI					x	x	x		x		x					**5**
public awareness, education about AI and its risks			x		x					x	x				x	**5**
future of employment					x					x	x				x	**4**
human autonomy	x		x	x					x							**4**
protection of whistleblowers							x									**1**
hidden costs (labeling, clickwork, material resources etc.)							x									**1**
affiliation (government, industry, science)	government	science	science	science	science	science	science	science	science	industry	industry	industry	industry	industry	industry	
number of ethical aspects	**8**	**7**	**11**	**11**	**11**	**9**	**11**	**5**	**11**	**10**	**8**	**6**	**6**	**5**	**8**	

Guidelines



privacy protection	14
accountability	13
fairness, non-discrimination, justice	13
transparency, openness	10
safety, cybersecurity	10
common good, sustainability	9
explainability, interpretabiliy	8
human oversight, control, auditing	8
dual-use problem, military, AI arms race	6
solidarity, inclusion, social cohesion	6
science-policy link	5
field-specific deliberations (health, military, mobility etc.)	5
diversity in the field of AI	5
public awareness, education about AI and its risks	5
future of employment	4
human autonomy	4
protection of whistleblowers	1
hidden costs (labeling, clickwork, material resources etc.)	1

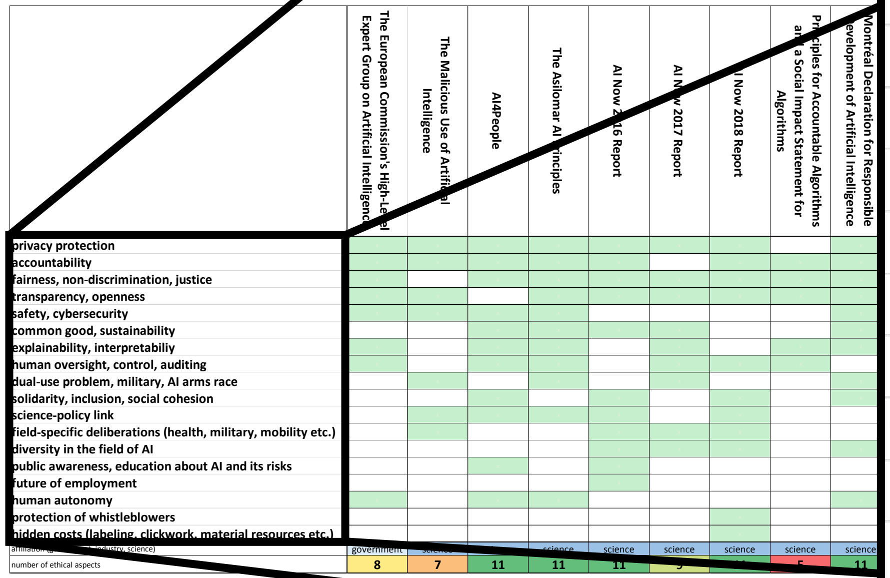
Guidelines

guidelines	The European Commission's High-Level Expert Group on Artificial Intelligence	The Malicious Use of Artificial Intelligence	AI4People	The Asilomar AI Principles	AI Now 2016 Report	AI Now 2017 Report	AI Now 2018 Report	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Montréal Declaration for Responsible Development of Artificial Intelligence	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems	ITI AI Policy Principles	Microsoft AI principles	Artificial Intelligence at Google	Everyday Ethics for Artificial Intelligence	Partnership on AI
notes on technical implementations	yes, but very few and superficial	yes, relatively comprehensive	none	none	none	none	none	none	none	yes, but very few and superficial	none	none	none	none	none
proportion of women among authors (f/m)	(8/10)	(5/21)	(5/8)	ns	(4/2)	(3/1)	(6/4)	(1/12)	(8/10)	ns	ns	ns	ns	(1/2)	ns
length (number of words)	16,546	34,017	8,609	646	11,530	18,273	25,759	13.59	4,754	40,915	2,272	75	882	4,488	1,481
affiliation (government, industry, science)	government	science	science	science	science	science	science	science	science	industry	industry	industry	industry	industry	industry
number of ethical aspects	8	7	11	11	11	9	11	5	11	10	8	6	6	5	8

accountability	13
fairness, non-discrimination, justice	13
transparency, openness	10
safety, cybersecurity	10
common good, sustainability	9
explainability, interpretabiliy	8
human oversight, control, auditing	8
dual-use problem, military, AI arms race	6
solidarity, inclusion, social cohesion	6
science-policy link	5
field-specific deliberations (health, military, mobility etc.)	5
diversity in the field of AI	5
public awareness, education about AI and its risks	5
future of employment	4
human autonomy	4
protection of whistleblowers	1
hidden costs (labeling, clickwork, material resources etc.)	1

affiliation (government, industry, science)	government	science	science	science	science	science	science	science	science	industry	industry	industry	industry	industry	industry
number of ethical aspects	8	7	11	11	11	9	11	5	11	10	8	6	6	5	8

Guidelines



privacy protection	14
accountability	13
fairness, non-discrimination, justice	13
transparency, openness	10
safety, cybersecurity	10
common good, sustainability	9
explainability, interpretabiliy	8
human oversight, control, auditing	8
dual-use problem, military, AI arms race	6
solidarity, inclusion, social cohesion	6
science-policy link	5
field-specific deliberations (health, military, mobility etc.)	5
diversity in the field of AI	5
public awareness, education about AI and its risks	5
future of employment	4
human autonomy	4
protection of whistleblowers	1
hidden costs (labeling, clickwork, material resources etc.)	1

Privacy (14/15)

- personality analysis, image recognition, disease prediction etc.

PNAS

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski[a,1], David Stillwell[a], and Thore Graepel[b]

[a]Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and [b]Microsoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

Computer-based personality judgments are more accurate than those made by humans

Wu Youyou[a,1,2], Michal Kosinski[b,1], and David Stillwell[a]

[a]Department of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom; and [b]Department of Computer Science, Stanford University, Stanford, CA 94305

Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)

Judging others' personalities is an essential skill in successful social living, as personality is a key driver behind people's interactions, behaviors, and emotions. Although accurate personality judgments stem from social-cognitive skills, developments in machine learning show that computer models can also make valid judgments. This study compares the accuracy of human and computer-based personality judgments, using a sample of 86,220 volunteers who completed a 100-item personality questionnaire. We show that (i) computer predictions based on a generic digital footprint (Facebook Likes) are more accurate ($r = 0.56$) than those made by the participants' Facebook friends using a personality questionnaire ($r = 0.49$); (ii) computer models show higher interjudge agreement; and (iii) computer personality judgments have higher external validity when predicting life outcomes such as substance use, political attitudes, and physical health; for some outcomes, they even outperform the self-rated personality scores. Computers outpacing humans in personality judgment presents significant opportunities and challenges in the areas of psychological assessment, marketing, and privacy.

personality judgment | social media | computational social science | artificial intelligence | big data

psychological traits (11). We used LASSO (least absolute shrinkage and selection operator) linear regressions (15) with 10-fold cross-validations, so that judgments for each participant were made using models developed on a different subsample of participants and their Likes. Likes are used by Facebook users to express positive association with online and offline objects, such as products, activities, sports, musicians, books, restaurants, or websites. Given the variety of objects, subjects, brands, and people that can be liked and the number of Facebook users (>1.3 billion), Likes represent one of the most generic kinds of digital footprint. For instance, liking a brand or a product offers a proxy for consumer preferences and purchasing behavior; music-related Likes reveal music taste; and liked websites allow for approximating web browsing behavior. Consequently, Like-based models offer a good proxy of what could be achieved based on a wide range of other digital footprints such as web browsing logs, web search queries, or purchase records (11).

Human personality judgments were obtained from the participants' Facebook friends, who were asked to describe a given

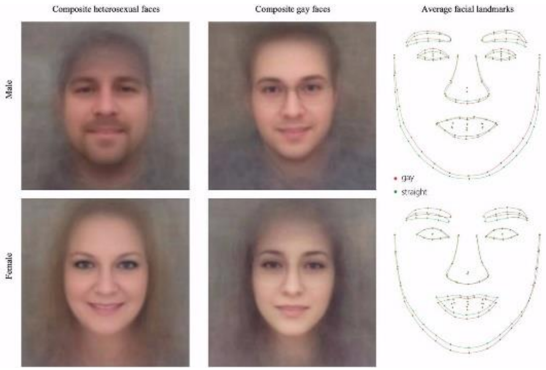
Significance

THIS IS A PREPRINT OF THE PEER REVIEWED ARTICLE TO APPEAR IN JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY.

THE MOST RECENT VERSION IS AVAILABLE AT https://osf.io/zn79k/

AUTHOR NOTES ARE AVAILABLE AT: https://goo.gl/9b2aR2

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images



Graduate School c

The study h

PNAS

Facebook language predicts depression in medical records

Johannes C. Eichstaedt[a,1,2], Robert J. Smith[b,1], Raina M. Merchant[b,c], Lyle H. Ungar[a,b], Patrick Crutchley[a,b], Daniel Preoţiuc-Pietro[a], David A. Asch[b,d], and H. Andrew Schwartz[e]

Journal of Alzheimer's Disease 49 (2016) 407–422
DOI 10.3233/JAD-150520
IOS Press

Linguistic Features Identify Alzheimer's Disease in Narrative Speech

Kathleen C. Fraser[a], Jed A. Meltzer[b] and Frank Rudzicz[a,c,*]
[a]Department of Computer Science, University of Toronto, Toronto, Canada
[b]Rotman Research Institute, Toronto, Canada
[c]Toronto Rehabilitation Institute-UHN, Toronto, Canada

Handling Associate Editor: Peter Garrard

Accepted 20 August 2015

Abstract.
Background: Although memory impairment is the main symptom of Alzheimer's disease (AD), language impairment can be an important marker. Relatively few studies of language in AD quantify the impairments in connected speech using computational techniques.
Objective: We aim to demonstrate state-of-the-art accuracy in automatically identifying Alzheimer's disease from short narrative samples elicited with a picture description task, and to uncover the salient linguistic factors with a statistical factor analysis.
Methods: Data are derived from the DementiaBank corpus, from which 167 patients diagnosed with "possible" or "probable" AD provide 240 narrative samples, and 97 controls provide an additional 233. We compute a number of linguistic variables from the transcripts, and acoustic variables from the associated audio files, and use these variables to train a machine learning classifier to distinguish between participants with AD and healthy controls. To examine the degree of heterogeneity of linguistic impairments in AD, we follow an exploratory factor analysis on these measures of speech and language with an oblique promax rotation, and provide interpretation for the resulting factors.
Results: We obtain state-of-the-art classification accuracies of over 81% in distinguishing individuals with AD from those without based on short samples of their language on a picture description task. Four clear factors emerge: semantic impairment, acoustic abnormality, syntactic impairment, and information impairment.
Conclusion: Modern machine learning and linguistic analysis will be increasingly useful in assessment and clustering of suspected AD.

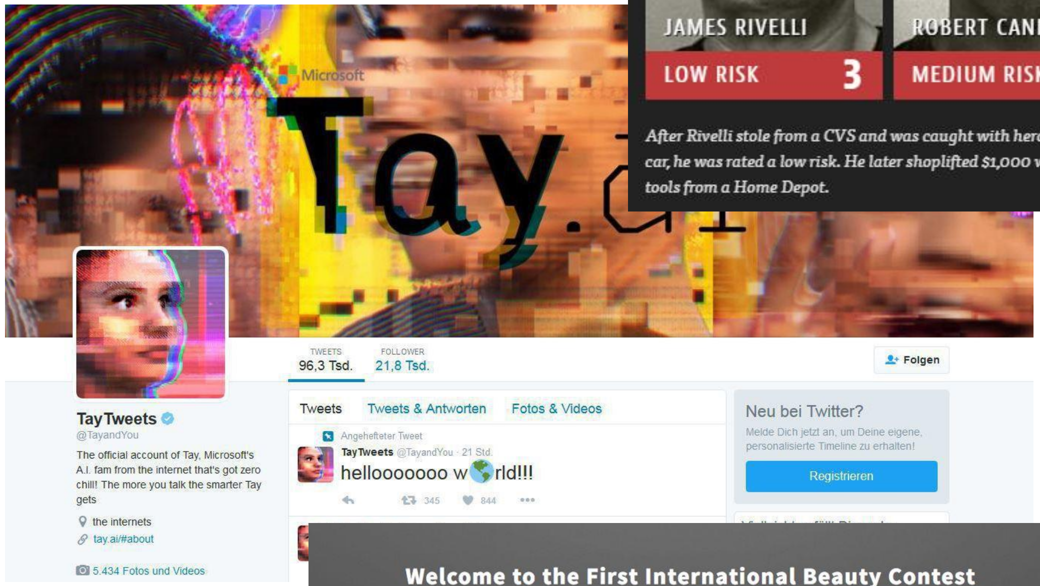
Keywords: Automatic data processing, factor analysis, geriatric assessment, heterogeneity, language, statistical

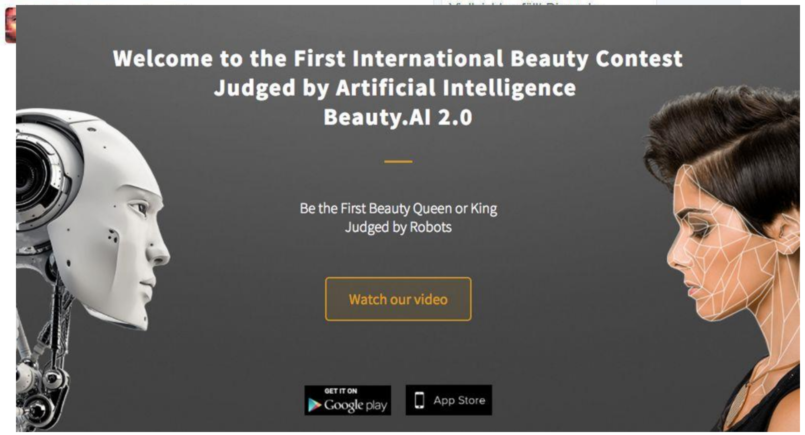
Accountability (13/15)

- who can be held legally responsible?
- AI systems as "e-persons"

Fairness (13/15)

- algorithmic discrimination
- bias in training data
- solutions provided by FAT ML community



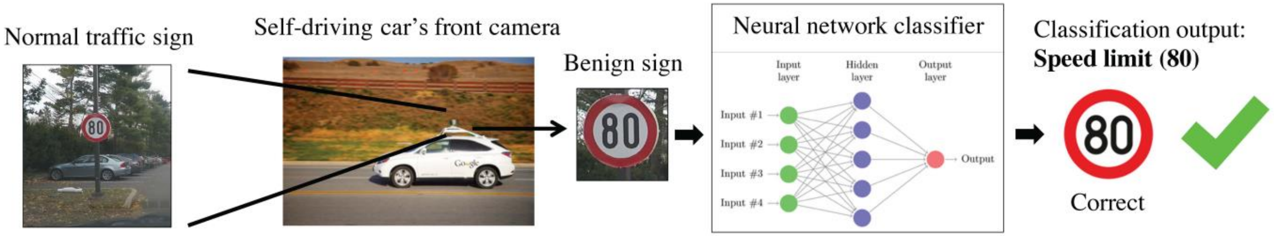


Transparency (10/15)

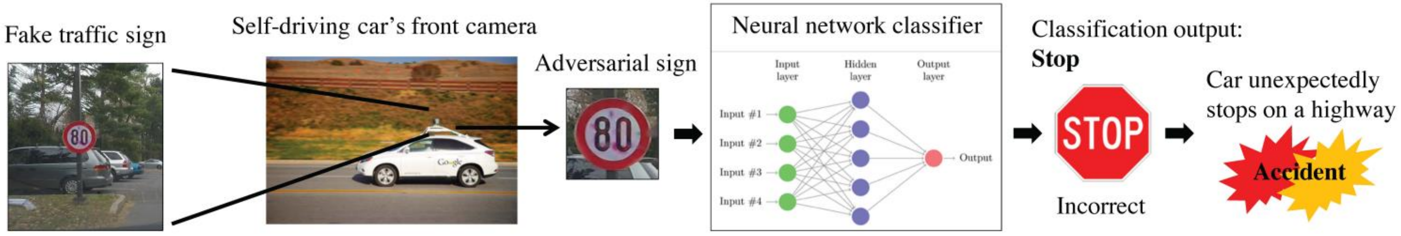
- problem of non-transparent organizations dealing with AI
- information asymmetries

Safety (10/15)

- dealing with security vulnerabilities
- data poisoning attacks, adversarial examples etc.



(a) Operation of the computer vision subsystem of an AV under *benign conditions*



(b) Operation of the computer vision subsystem of an AV under *adversarial conditions*

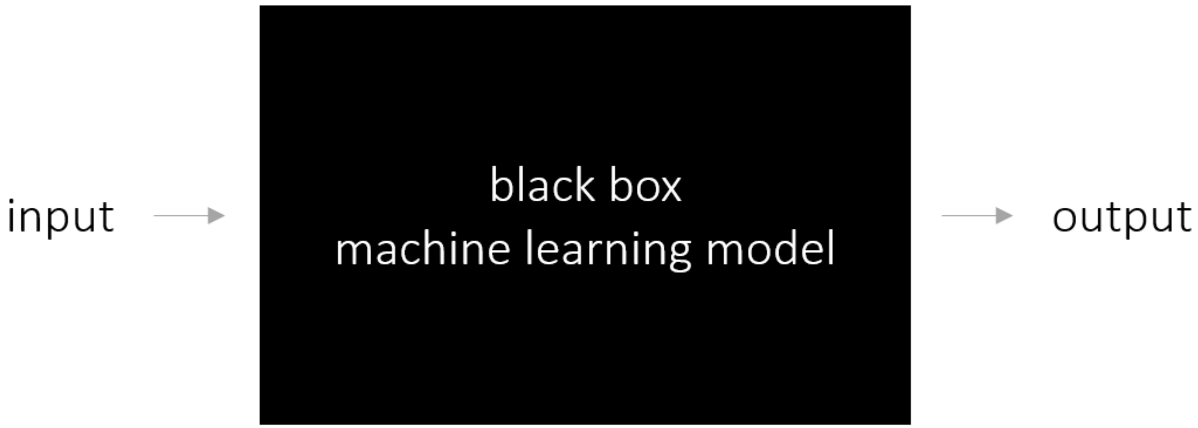
Common good (9/15)

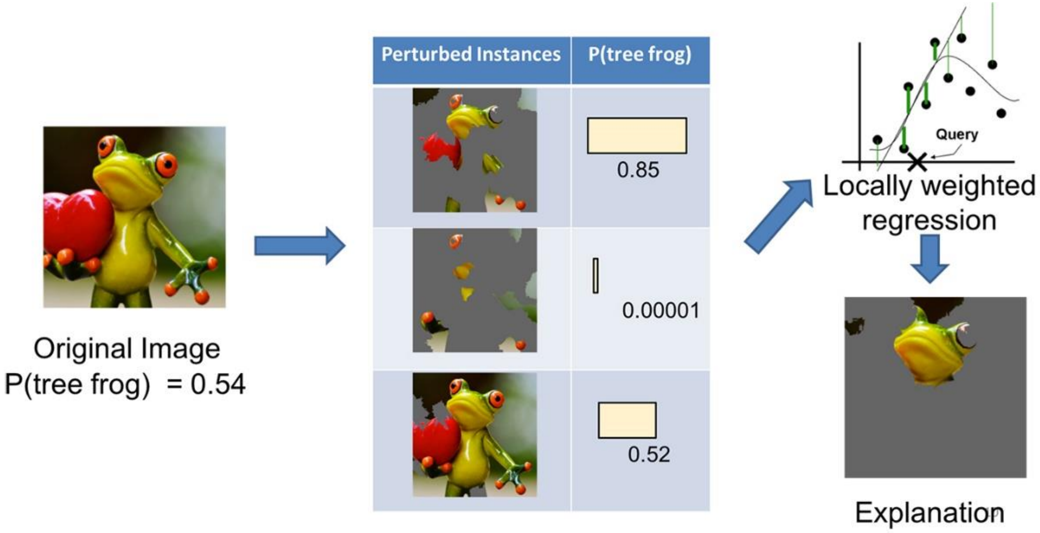
- idea of AI fostering sustainability goals
- AI4Good, Beneficial AI etc.

AI for Good
FOUNDATION

Explainability (8/15)

- black box problems
- XAI



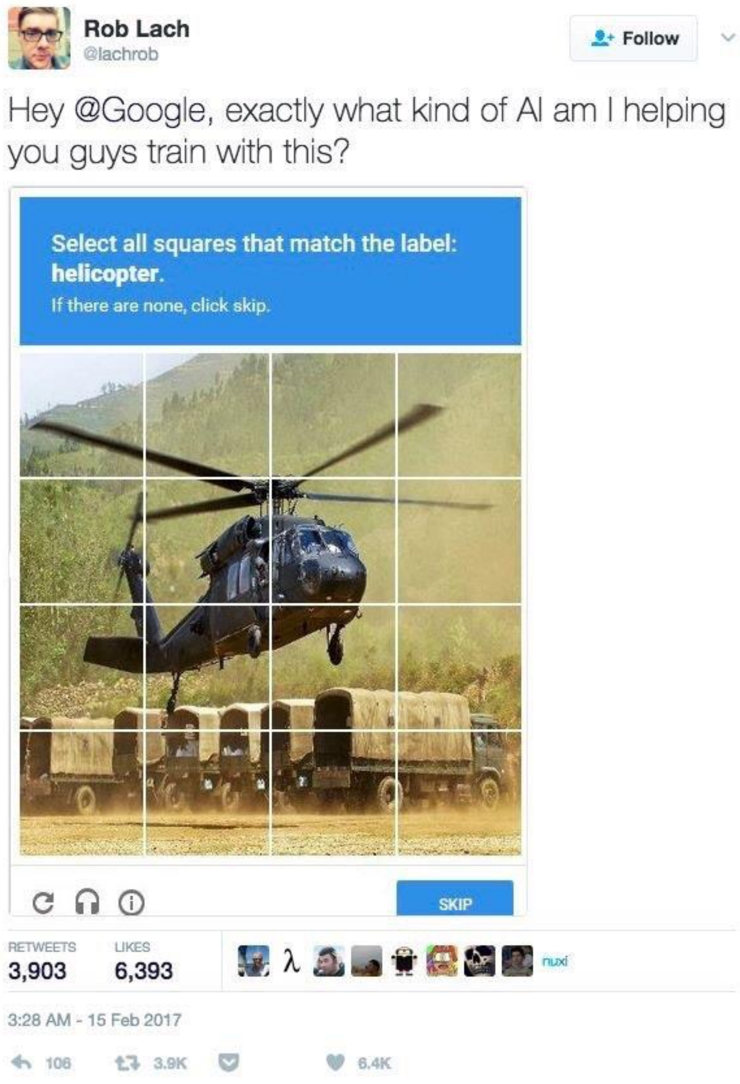


Human oversight (8/15)

- developing auditing mechanisms
- human in the loop

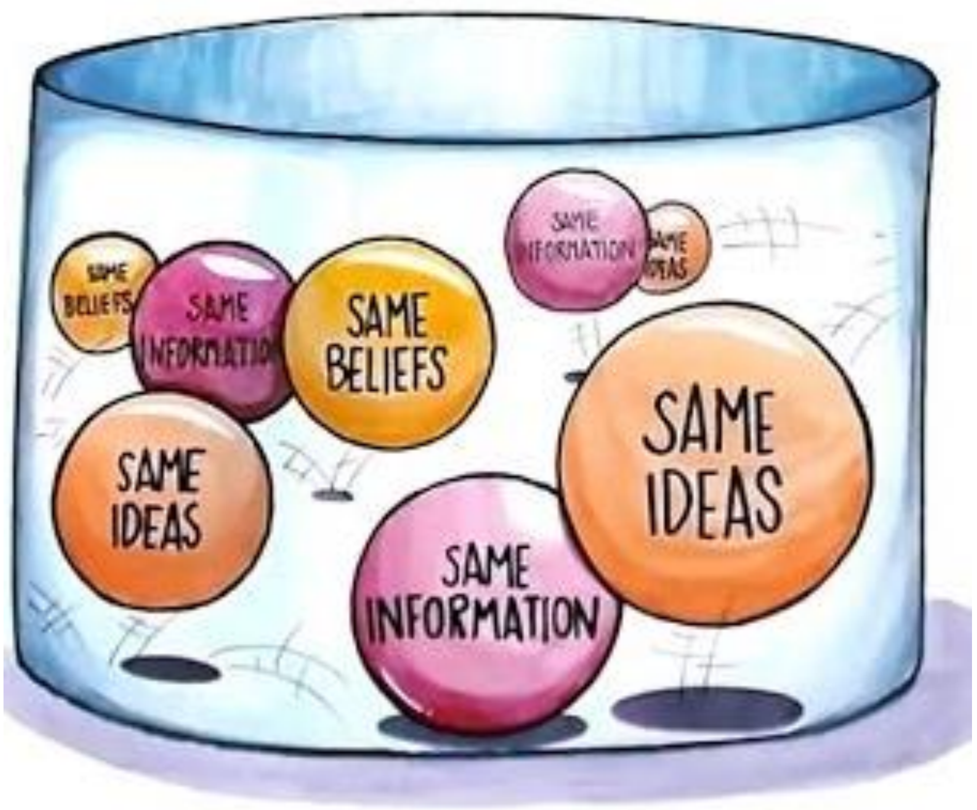
Dual-use problem (06/15)

- machine learning as “general purpose technology”
- opposing the military use of AI



Solidarity, social cohesion (6/15)

- AI and social media
- speaking against filter bubbles, micro targeting, radicalization etc.



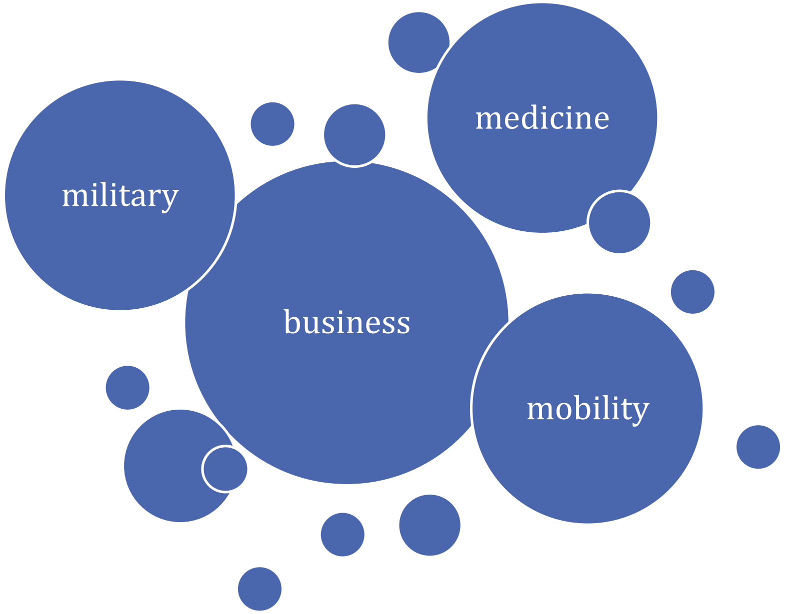
Science-policy link (5/15)

- multistakeholder approach
- connecting science, industry and politics

4 PEOPLE

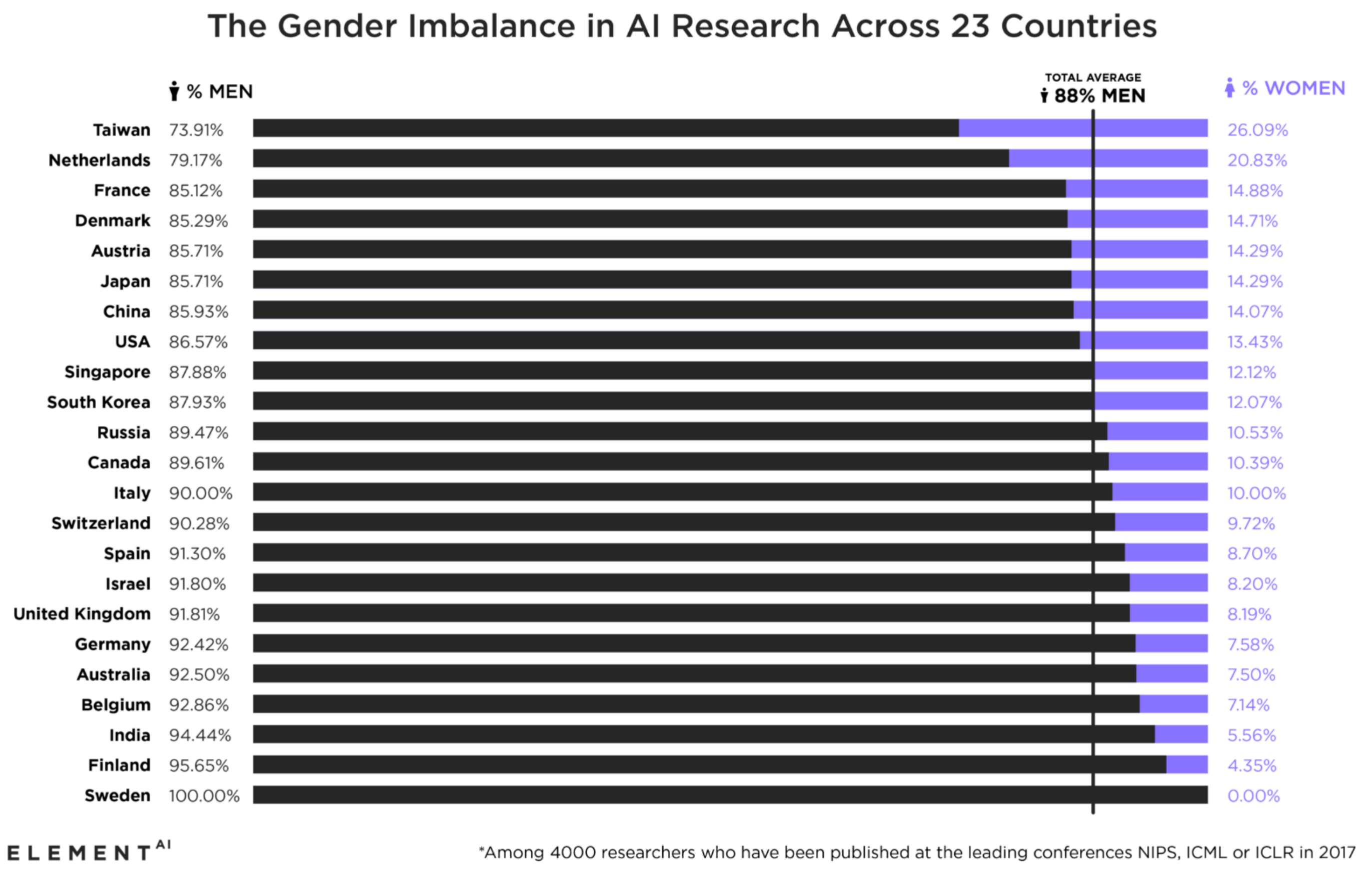
Field-specific deliberations (5/15)

- AI in specific social systems or fields
- medicine, military, mobility etc.



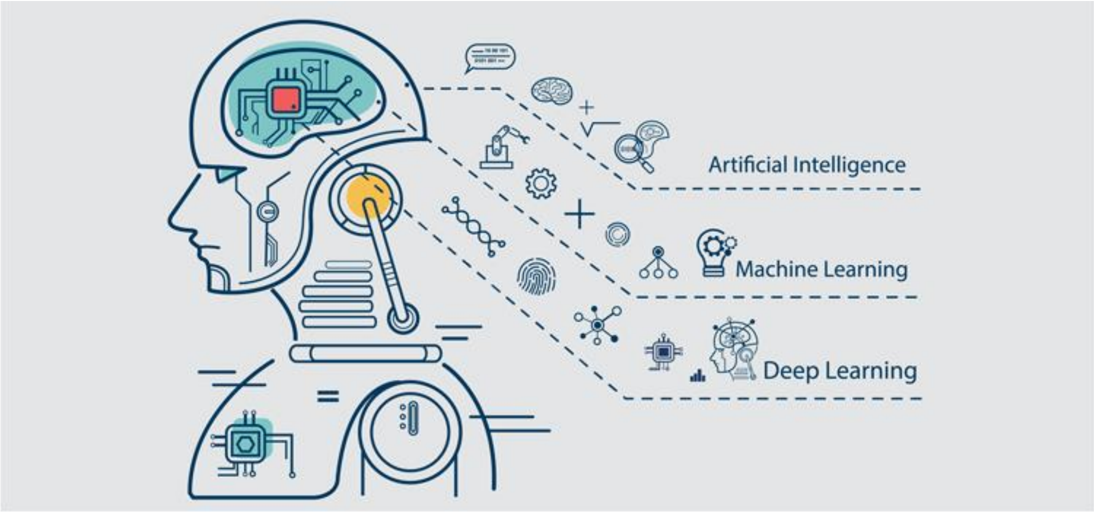
Diversity in the field of AI (5/15)

- diversity crisis in the AI sector
- statistics show blatant inequalities



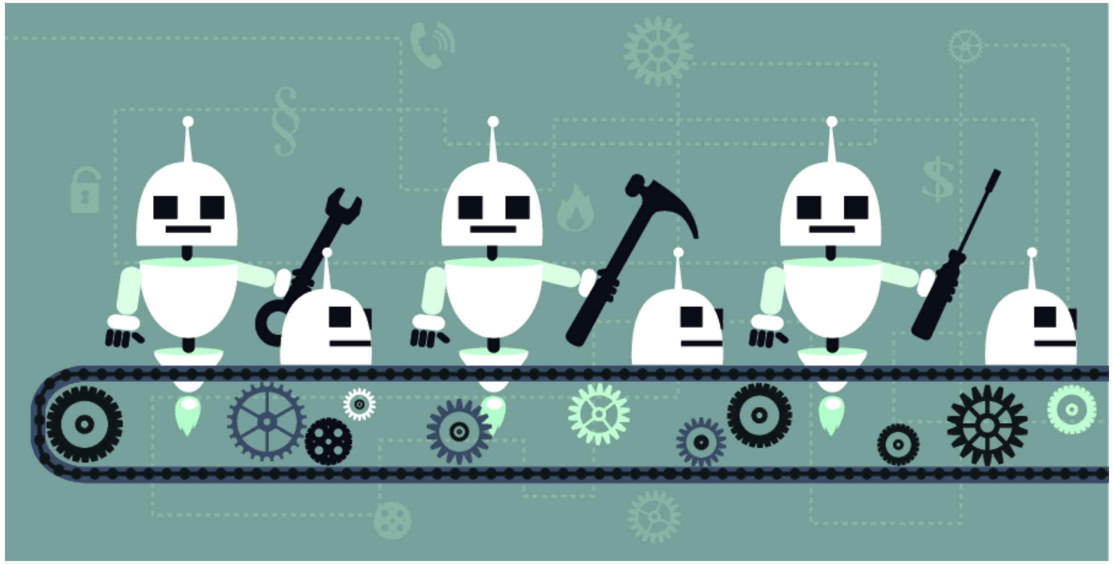
Public awareness, education about AI (5/15)

- creation of educational curricula and public awareness activities



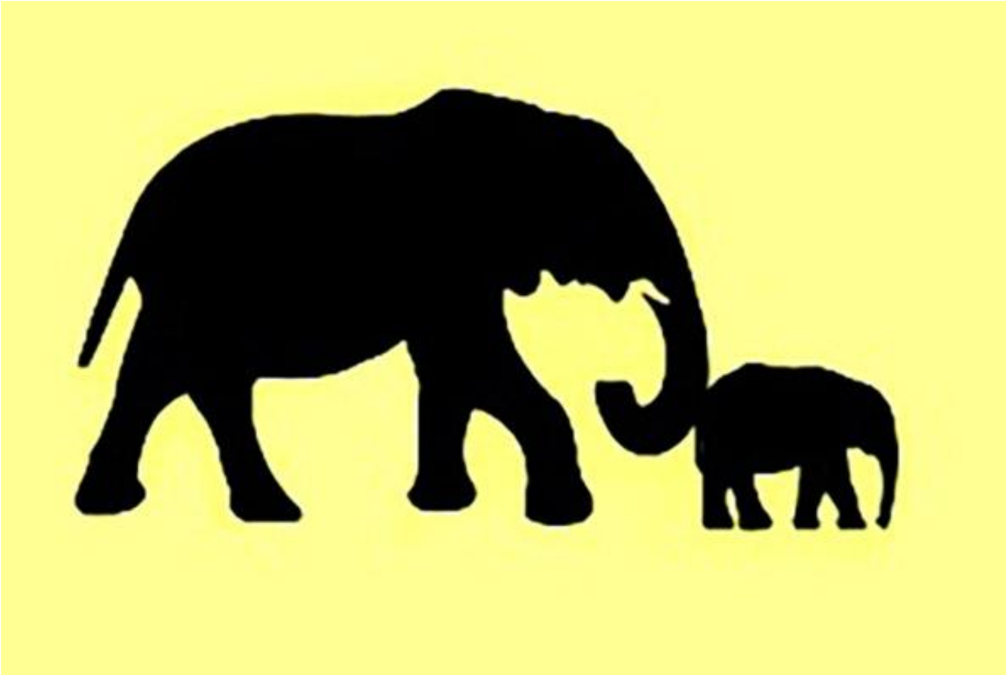
Future of employment (4/15)

- ideas about robot taxes, universal basic income etc.



Human autonomy (4/15)

- not using AI for behavior manipulation
- nudging, micro targeting, personalized online advertising, captology, etc.

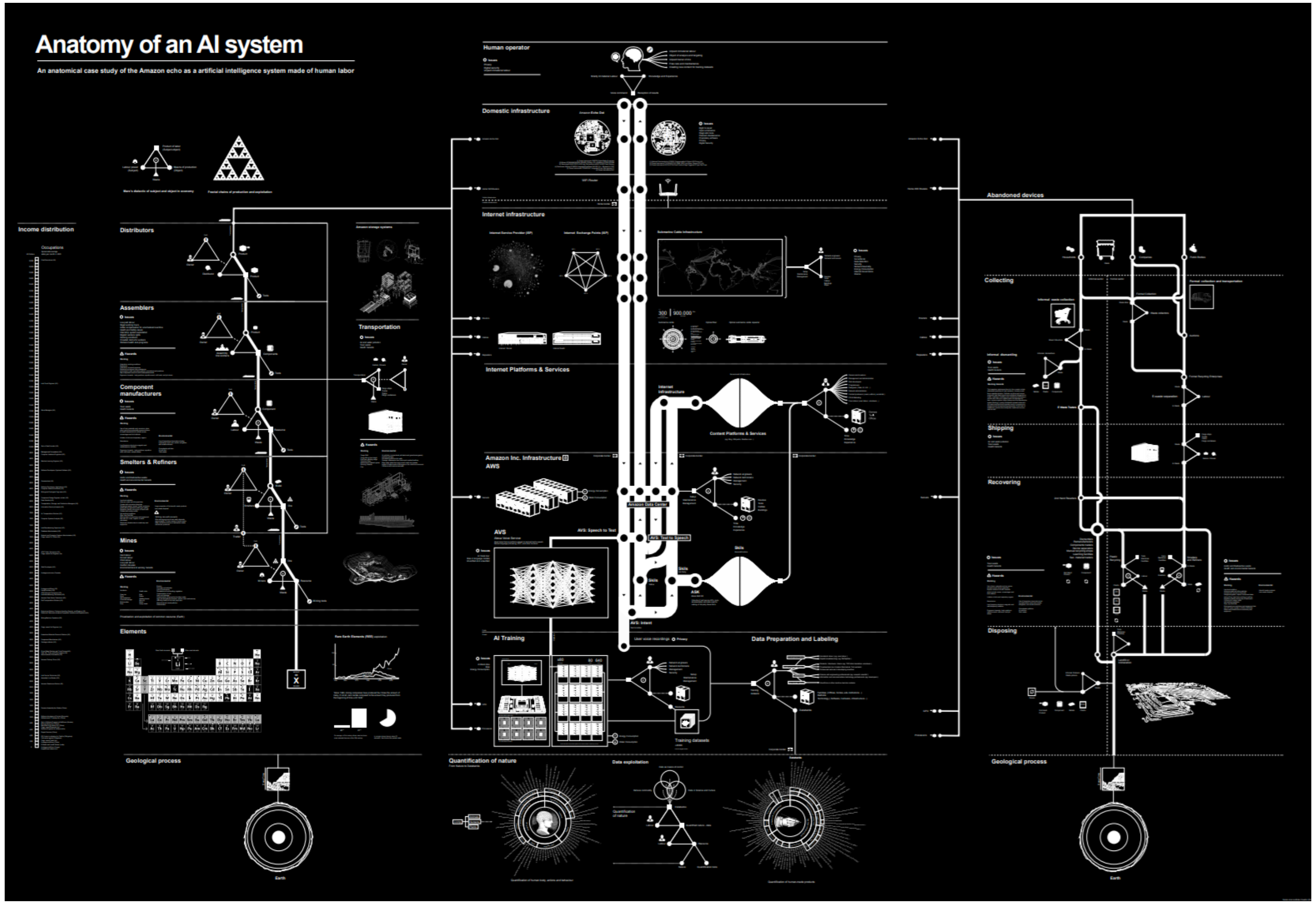


Protection of whistleblowers (2/15)

- need for better protection

Hidden costs (1/15)

- labeling factories (clickwork), content moderation, energy, material resources etc.



Guidelines

	The European Commission's High-Level Expert Group on Artificial Intelligence	The Malicious Use of Artificial Intelligence	AI4People	The Asilomar AI Principles	AI Now 2016 Report	AI Now 2017 Report	AI Now 2018 Report	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Montréal Declaration for Responsible Development of Artificial Intelligence	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems	ITI AI Policy Principles	Microsoft AI principles	Artificial Intelligence at Google	Everyday Ethics for Artificial Intelligence	Partnership on AI	number of mentions
privacy protection	x	x	x	x	x	x	x		x	x	x	x	x	x	x	**14**
accountability	x	x	x	x	x		x	x	x	x	x	x		x	x	**13**
fairness, non-discrimination, justice	x		x	x	x	x	x	x	x	x		x	x	x	x	**13**
transparency, openness	x	x		x	x	x	x	x	x	x		x				**10**
safety, cybersecurity	x	x	x	x					x	x	x	x	x		x	**10**
common good, sustainability			x	x	x	x			x	x			x	x	x	**9**
explainability, interpretabiliy	x		x	x		x		x	x		x			x		**8**
human oversight, control, auditing	x		x	x		x	x	x			x		x			**8**
dual-use problem, military, AI arms race		x		x		x			x	x			x			**6**
solidarity, inclusion, social cohesion			x		x		x		x			x			x	**6**
science-policy link		x	x	x	x		x									**5**
field-specific deliberations (health, military, mobility etc.)		x			x	x	x			x						**5**
diversity in the field of AI					x	x	x		x		x					**5**
public awareness, education about AI and its risks			x		x					x	x				x	**5**
future of employment					x					x	x				x	**4**
human autonomy	x		x	x					x							**4**
protection of whistleblowers							x									**1**
hidden costs (labeling, clickwork, material resources etc.)							x									**1**
affiliation (government, industry, science)	government	science	science	science	science	science	science	science	science	industry	industry	industry	industry	industry	industry	
number of ethical aspects	**8**	**7**	**11**	**11**	**11**	**9**	**11**	**5**	**11**	**10**	**8**	**6**	**6**	**5**	**8**	

„Unethical“ behaviour (Kish-Gephart et al. 2010)

individual characteristics

moral development
idealism
machiavellianism
job satisfaction
education level

characteristics of the moral issue

concentration of consequences
magnitude of consequences
probability of effect
proximity

environment characteristics

egoistic ethical climate
benevolent ethical climate
code existence
code enforcement
appreciation of ethical behaviour

„unethical“ behaviour / intentions in organisations

Dr. Thilo Hagendorff

University of Tuebingen
Cluster of Excellence Machine Learning

thilo.hagendorff@uni-tuebingen.de
www.thilo-hagendorff.info