KI-Tugenden

Ein neuer Zugang zur praktischen Umsetzung von KI-Ethik

Dr. Thilo Hagendorff

Universität Tübingen

Exzellenzcluster "Maschinelles Lernen: Neue Perspektiven für die Wissenschaft"

Internationales Zentrum für Ethik in den Wissenschaften, "Forum Privatheit"

19.11.2021

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

forum <privatheit> selbstbestimmtes_leben_in_der_digitalen_welt

Kontext

- Was ist das Ziel?
  - vertrauenswürdige, menschzentrierte, sichere, „beneficial“ KI
- Wie?

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Kritische Stimmen

- Forderungen nach effektiver KI -Ethik
- Wie erreichen?

6 Reasons Why AI Ethics in Corporations is All Talk and No Action

Mia Dand Follow
Sep 28, 2018 · 6 min read

source: https://becominghuman.ai/6-reasons-why-ai-ethics-in-corporations-is-all-talk-and-no-action-4f126af42668

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit> selbstbestimmtes leben in der digitalen welt

Gegenwärtiger Zugang

- Prinzipienbasierter, deontologischer Zugang
- umfassende Guidelines
- Guidelines aus Wissenschaft, Industrie, oder staatlichen Kontexten

Minds and Machines (2018) 28:689–707
https://doi.org/10.1007/s11023-018-9482-5

CrossMark

AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations

Luciano Floridi[1,2] · Josh Cowls[1,2] · Monica Beltrametti[3] · Raja Chatila[4,5] · Patrice Chazerand[6] · Virginia Dignum[7,8] · Christoph Luetge[9] · Robert Madelin[10] · Ugo Pagallo[11] · Francesca Rossi[12,13] · Burkhard Schafer[14] · Peggy Valcke[15,16] · Effy Vayena[17]

Received: 28 October 2018 / Accepted: 2 November 2018 / Published online: 26 November 2018
© The Author(s) 2018

Abstract
This article reports the findings of AI4People, an Atomium—EISMD initiative designed to lay the foundations for a "Good AI Society". We introduce the core opportunities and risks of AI for society; present a synthesis of five ethical principles that should undergird its development and adoption; and offer 20 concrete recommendations—to assess, to develop, to incentivise, and to support good AI—which in some cases may be undertaken directly by national or supranational policy makers, while in others may be led by other stakeholders. If adopted, these recommendations would serve as a firm foundation for the establishment of a Good AI Socie

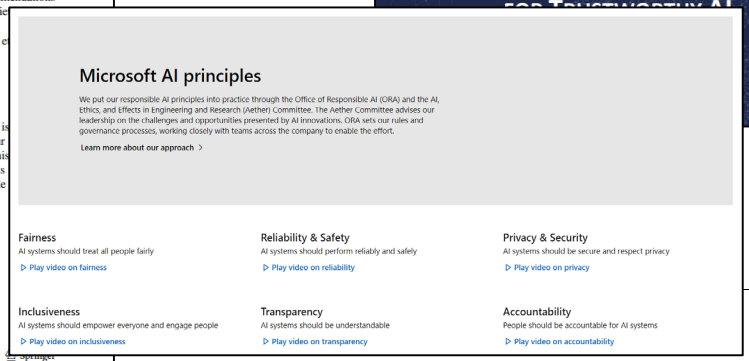
Keywords Artificial intelligence · AI4People · Data governance · Digital e
Governance · Ethics of AI

1 Introduction

AI is not another utility that needs to be regulated once it is mature. It is
ful force, a new form of smart agency, which is already reshaping our
interactions, and our environments. AI4People was set up to help steer this
force towards the good of society, everyone in it, and the environments
This article is the outcome of the collaborative effort by the AI4People

Luciano Floridi
luciano.floridi@oii.ox.ac.uk

Extended author information available on the last page of the article

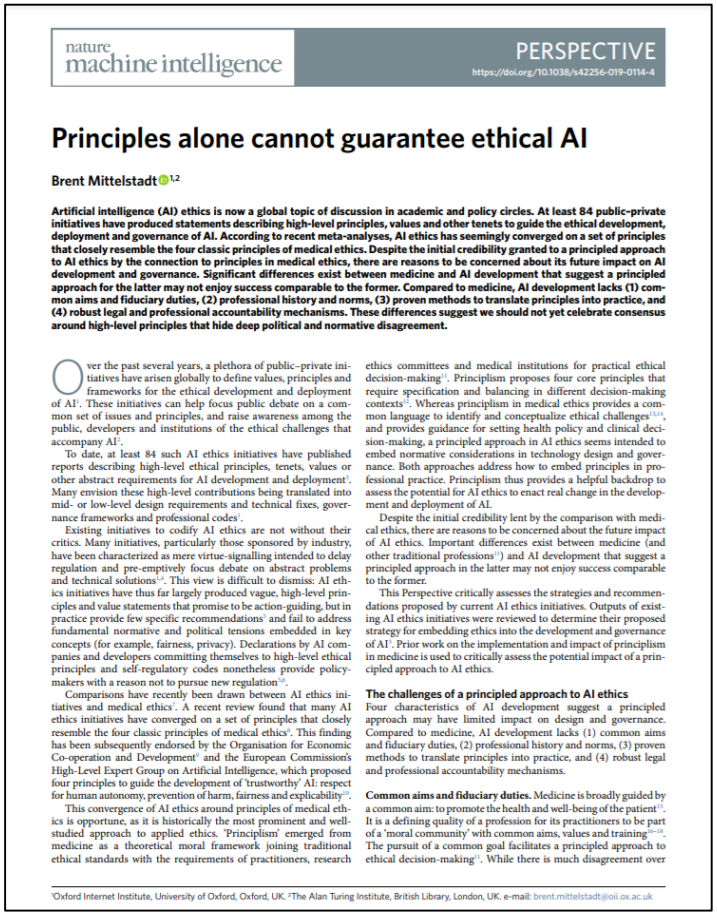


GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit>
selbstbestimmtes_leben_in_der_digitalen_welt

Kritik

- Ethik als Marketingstrategie
- KI-Ethik ist ineffizient
- Prinzipien entbehren Praktikabilität und sind zu abstrakt



nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-019-0114-4

Principles alone cannot guarantee ethical AI

Brent Mittelstadt[1,2]

Artificial intelligence (AI) ethics is now a global topic of discussion in academic and policy circles. At least 84 public–private initiatives have produced statements describing high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI. According to recent meta-analyses, AI ethics has seemingly converged on a set of principles that closely resemble the four classic principles of medical ethics. Despite the initial credibility granted to a principled approach to AI ethics by the connection to principles in medical ethics, there are reasons to be concerned about its future impact on AI development and governance. Significant differences exist between medicine and AI development that suggest a principled approach for the latter may not enjoy success comparable to the former. Compared to medicine, AI development lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms. These differences suggest we should not yet celebrate consensus around high-level principles that hide deep political and normative disagreement.

Over the past several years, a plethora of public–private initiatives have arisen globally to define values, principles and frameworks for the ethical development and deployment of AI[1]. These initiatives can help focus public debate on a common set of issues and principles, and raise awareness among the public, developers and institutions of the ethical challenges that accompany AI[2].

To date, at least 84 such AI ethics initiatives have published reports describing high-level ethical principles, tenets, values or other abstract requirements for AI development and deployment[3]. Many envision these high-level contributions being translated into mid- or low-level design requirements and technical fixes, governance frameworks and professional codes[3].

Existing initiatives to codify AI ethics are not without their critics. Many initiatives, particularly those sponsored by industry, have been characterized as mere virtue-signalling intended to delay regulation and pre-emptively focus debate on abstract problems and technical solutions[4,5]. This view is difficult to dismiss: AI ethics initiatives have thus far largely produced vague, high-level principles and value statements that promise to be action-guiding, but in practice provide few specific recommendations[3] and fail to address fundamental normative and political tensions embedded in key concepts (for example, fairness, privacy). Declarations by AI companies and developers committing themselves to high-level ethical principles and self-regulatory codes nonetheless provide policymakers with a reason not to pursue new regulation[5,6].

Comparisons have recently been drawn between AI ethics initiatives and medical ethics[3]. A recent review found that many AI ethics initiatives have converged on a set of principles that closely resemble the four classic principles of medical ethics[6]. This finding has been subsequently endorsed by the Organisation for Economic Co-operation and Development[7] and the European Commission's High-Level Expert Group on Artificial Intelligence, which proposed four principles to guide the development of 'trustworthy' AI: respect for human autonomy, prevention of harm, fairness and explicability[8,9].

This convergence of AI ethics around principles of medical ethics is opportune, as it is historically the most prominent and well-studied approach to applied ethics. 'Principlism' emerged from medicine as a theoretical moral framework joining traditional ethical standards with the requirements of practitioners, research ethics committees and medical institutions for practical ethical decision-making[11]. Principlism proposes four core principles that require specification and balancing in different decision-making contexts[12]. Whereas principlism in medical ethics provides a common language to identify and conceptualize ethical challenges[13,14], and provides guidance for setting health policy and clinical decision-making, a principled approach in AI ethics seems intended to embed normative considerations in technology design and governance. Both approaches address how to embed principles in professional practice. Principlism thus provides a helpful backdrop to assess the potential for AI ethics to enact real change in the development and deployment of AI.

Despite the initial credibility lent by the comparison with medical ethics, there are reasons to be concerned about the future impact of AI ethics. Important differences exist between medicine (and other traditional professions[15]) and AI development that suggest a principled approach in the latter may not enjoy success comparable to the former.

This Perspective critically assesses the strategies and recommendations proposed by current AI ethics initiatives. Outputs of existing AI ethics initiatives were reviewed to determine their proposed strategy for embedding ethics into the development and governance of AI[3]. Prior work on the implementation and impact of principlism in medicine is used to critically assess the potential impact of a principled approach to AI ethics.

The challenges of a principled approach to AI ethics

Four characteristics of AI development suggest a principled approach may have limited impact on design and governance. Compared to medicine, AI development lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms.

Common aims and fiduciary duties. Medicine is broadly guided by a common aim: to promote the health and well-being of the patient[15]. It is a defining quality of a profession for its practitioners to be part of a 'moral community' with common aims, values and training[16–18]. The pursuit of a common goal facilitates a principled approach to ethical decision-making[15]. While there is much disagreement over

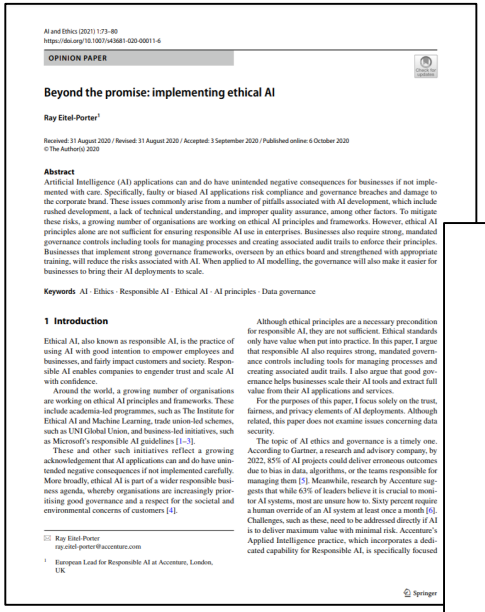
[1]Oxford Internet Institute, University of Oxford, Oxford, UK. [2]The Alan Turing Institute, British Library, London, UK. e-mail: brent.mittelstadt@oii.ox.ac.uk

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

forum <privatheit> selbstbestimmtes_leben_in_der_digitalen_welt

Von Prinzipien zur Praxis



AI and Ethics (2021) 1:73–80
https://doi.org/10.1007/s43681-020-00011-6

OPINION PAPER

Beyond the promise: implementing ethical AI

Ray Eitel-Porter[1]

Received: 31 August 2020 / Revised: 31 August 2020 / Accepted: 3 September 2020 / Published online: 6 October 2020
© The Author(s) 2020

Abstract
Artificial Intelligence (AI) applications can and do have unintended negative consequences for businesses if not implemented with care. Specifically, faulty or biased AI applications risk compliance and governance breaches and damage to the corporate brand. These issues commonly arise from a number of pitfalls associated with AI development, which include rushed development, a lack of technical understanding, and improper quality assurance, among other factors. To mitigate these risks, a growing number of organisations are working on ethical AI principles and frameworks. However, ethical AI principles alone are not sufficient for ensuring responsible AI use in enterprises. Businesses also require strong, mandated governance controls including tools for managing processes and creating associated audit trails to enforce their principles. Businesses that implement strong governance frameworks, overseen by an ethics board and strengthened with appropriate training, will reduce the risks associated with AI. When applied to AI modelling, the governance will also make it easier for businesses to bring their AI deployments to scale.

Keywords AI · Ethics · Responsible AI · Ethical AI · AI principles · Data governance

1 Introduction

Ethical AI, also known as responsible AI, is the practice of using AI with good intention to empower employees and businesses, and fairly impact customers and society. Responsible AI enables companies to engender trust and scale AI with confidence.

Around the world, a growing number of organisations are working on ethical AI principles and frameworks. These include academia-led programmes, such as The Institute for Ethical AI and Machine Learning, trade union-led schemes, such as UNI Global Union, and business-led initiatives, such as Microsoft's responsible AI guidelines [1–3].

These and other such initiatives reflect a growing acknowledgement that AI applications can and do have unintended negative consequences if not implemented carefully. More broadly, ethical AI is part of a wider responsible business agenda, whereby organisations are increasingly prioritising good governance and a respect for the societal and environmental concerns of customers [4].

Although ethical principles are a necessary precondition for responsible AI, they are not sufficient. Ethical standards only have value when put into practice. In this paper, I argue that responsible AI also requires strong, mandated governance controls including tools for managing processes and creating associated audit trails. I also argue that good governance helps businesses scale their AI tools and extract full value from their AI applications and services.

For the purposes of this paper, I focus solely on the trust, fairness, and privacy elements of AI deployments. Although related, this paper does not examine issues concerning data security.

The topic of AI ethics and governance is a timely one. According to Gartner, a research and advisory company, by 2022, 85% of AI projects could deliver erroneous outcomes due to bias in data, algorithms, or the teams responsible for managing them [5]. Meanwhile, research by Accenture suggests that while 63% of leaders believe it is crucial to monitor AI systems, most are unsure how to. Sixty percent require a human override of an AI system at least once a month [6]. Challenges, such as these, need to be addressed directly if AI is to deliver maximum value with minimal risk. Accenture's Applied Intelligence practice, which incorporates a dedicated capability for Responsible AI, is specifically focused

✉ Ray Eitel-Porter
ray.eitel-porter@accenture.com

[1] European Lead for Responsible AI at Accenture, London, UK

Science and Engineering Ethics (2020) 26:2141–2168
https://doi.org/10.1007/s11948-019-00165-5

ORIGINAL RESEARCH/SCHOLARSHIP

From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices

Jessica Morley[1] · Luciano Floridi[1,2] · Libby Kinsey[3] · Anat Elhalal[3]

Received: 16 May 2019 / Accepted: 29 November 2019 / Published online: 11 December 2019
© The Author(s) 2019

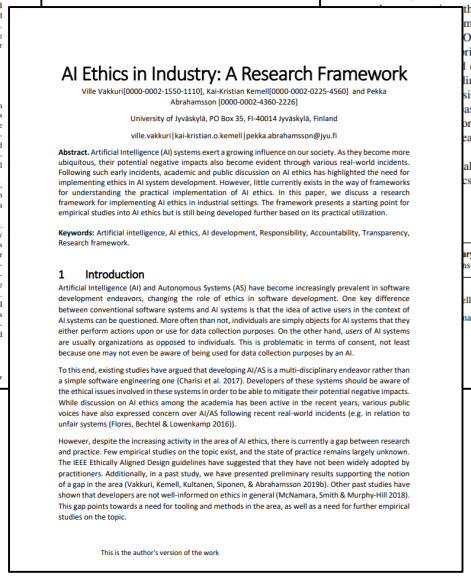
Abstract
The debate about the ethical implications of Artificial Intelligence dates from the 1960s (Samuel in Science, 132(3429):741–742, 1960. https://doi.org/10.1126/science.132.3429.741; Wiener in Cybernetics: or control and communication in the animal and the machine, MIT Press, New York, 1961). However, in recent ye
bolic AI has been complemented and sometimes replaced by (Deep) Neu
works and Machine Learning (ML) techniques. This has vastly increased its
utility and impact on society, with the consequence that the ethical debate
mainstream. Such a debate has primarily focused on principles—the 'wha
ethics (beneficence, non-maleficence, autonomy, justice and explicability)
he 'how.' Awareness of the potential issues is increasing
mmunity's ability to take action to mitigate the associated
Our intention in presenting this research is to contribute to
rinciples and practices by constructing a typology that n
developers apply ethics at each stage of the Machine L
line, and to signal to researchers where further work is
sively on Machine Learning, but it is hoped that the result
asily applicable to other branches of AI. The article outl
or creating this typology, the initial findings, and provides
earch needs.

al intelligence · Applied ethics · Data governance · Digital
cs of AI · Machine learning

ary material The online version of this article (https://doi.org/10.1007/s1
us supplementary material, which is available to authorized users.

ellogg.ox.ac.uk

nation available on the last page of the article



AI Ethics in Industry: A Research Framework

Ville Vakkuri[0000-0002-1550-1110], Kai-Kristian Kemell[0000-0002-0225-4560] and Pekka Abrahamsson [0000-0002-4360-2226]

University of Jyväskylä, PO Box 35, FI-40014 Jyväskylä, Finland

ville.vakkuri|kai-kristian.o.kemell|pekka.abrahamsson@jyu.fi

Abstract. Artificial Intelligence (AI) systems exert a growing influence on our society. As they become more ubiquitous, their potential negative impacts also become evident through various real-world incidents. Following such early incidents, academic and public discussion on AI ethics has highlighted the need for implementing ethics in AI system development. However, little currently exists in the way of frameworks for understanding the practical implementation of AI ethics. In this paper, we discuss a research framework for implementing AI ethics in industrial settings. The framework presents a starting point for empirical studies into AI ethics but is still being developed further based on its practical utilization.

Keywords: Artificial intelligence, AI ethics, AI development, Responsibility, Accountability, Transparency, Research framework.

1 Introduction

Artificial Intelligence (AI) and Autonomous Systems (AS) have become increasingly prevalent in software development endeavors, changing the role of ethics in software development. One key difference between conventional software systems and AI systems is that the idea of active users in the context of AI systems can be questioned. More often than not, individuals are simply objects for AI systems that they either perform actions upon or use for data collection purposes. On the other hand, users of AI systems are usually organizations as opposed to individuals. This is problematic in terms of consent, not least because one may not even be aware of being used for data collection purposes by an AI.

To this end, existing studies have argued that developing AI/AS is a multi-disciplinary endeavor rather than a simple software engineering one (Charisi et al. 2017). Developers of these systems should be aware of the ethical issues involved in these systems in order to be able to mitigate their potential negative impacts. While discussion on AI ethics among the academia has been active in the recent years, various public voices have also expressed concern over AI/AS following recent real-world incidents (e.g. in relation to unfair systems (Flores, Bechtel & Lowenkamp 2016)).

However, despite the increasing activity in the area of AI ethics, there is currently a gap between research and practice. Few empirical studies on the topic exist, and the state of practice remains largely unknown. The IEEE Ethically Aligned Design guidelines have suggested that they have not been widely adopted by practitioners. Additionally, in a past study, we have presented preliminary results supporting the notion of a gap in the area (Vakkuri, Kemell, Kultanen, Siponen, & Abrahamsson 2019b). Other past studies have shown that developers are not well-informed on ethics in general (McNamara, Smith & Murphy-Hill 2018). This gap points towards a need for tooling and methods in the area, as well as a need for further empirical studies on the topic.

comment

Towards ethical and socio-legal governance in AI

Many high-level ethics guidelines for AI have been produced in the past few years. It is time to work towards concrete policies within the context of existing moral, legal and cultural values, say Andreas Theodorou and Virginia Dignum.

Andreas Theodorou and Virginia Dignum

Artificial intelligence (AI) is increasingly believed to be the main transformative technology of our time. A pressing question is how to develop and deploy AI systems that are aligned with fundamental human principles and our legal system, and that serve the common good. As AI is now applied across many different domains, governments and policymakers are exploring how to shape the process of decision making and implementation of governmental policy in ways that ensure public safety, social stability and continued innovation. In this Comment, we discuss the challenges of producing such concrete policies, critique existing guidelines, and suggest practical steps forward at establishing a socially beneficial policy for AI through the use of technical standards, legislation and education.

Governance is necessary to reduce the number of negative incidents, ensure trust and create long-term societal stability through the use of well-established tools and design practices. Well-designed regulations do not eliminate innovation but instead enhance it through the development and promotion of both socio-legal and technical means to enforce compliance[1]. Moreover, policy is needed to determine human responsibility in the development and deployment of intelligent systems, filling the gap that emerges from the increased automation of decision. Furthermore, the ultimate aim of regulation is to ensure well-being for all in a sustainable world so it can guide responsible research, development and use of AI.

Ethical AI gives us more responsibility

The term AI refers to different issues, each of which has its own terminology, communities and expectations[2]. Depending on the focus and the context, AI can refer to:

1. A (computational) technology that is able to infer patterns and possibly draw conclusions from data; currently AI technologies are often based on machine learning and/or neural networking-based paradigms.
2. A field of scientific research (this is the original reference and still predominant in academia); the field of AI includes the study of theories and methods for adaptability, interaction and autonomy of machines (virtual or physical).
3. An (autonomous) entity (for example, when one refers to 'an' AI); this is the most common reference in media and science fiction but is the most inaccurate one — it often brings with it the (dystopic) view of malicious power.

Lack of understanding of these differences is one of the reasons why many AI guidelines and principles are so varied. Right now, this confusion leads to constant rewriting of similar high-level statements, creates exploitable loopholes and increases the public's misconceptions. Organizations can circumvent even their own AI ethics policies by claiming that a product is 'dumb' enough to avoid any need for compliance to any AI-specific policy or 'smart' enough to claim 'autonomy' and pass any responsibility to the machine. For example, after the insolvency of Air Berlin, Lufthansa's prices for flights within Germany increased up to 30%[3]. Lufthansa's response to an enquiry from the German consumer protection organization was that the algorithm acts autonomously and beyond the company's direct control. At the time of writing, Apple and Goldman Sachs are under investigation by New York's Department of Financial Services for discriminatory decisions made by the application system used in their newly launched credit card; male applicants for the card received significantly higher credit limit than their female spouses, even if the latter had higher credit scores[4]. When applicants tried to appeal these credit limits, the response was that the algorithm makes the final decision and its actual decision-making system is a black box that cannot be challenged.

It is fundamental to recognize that technology, or the artefact that embeds that technology, cannot be separated from the socio-technical system of which it is a component. This system includes people and organizations in many different roles (for example, developer, manufacturer, user, bystander or policymaker), their interactions and the processes that organize these interactions. Guidelines, principles and strategies must direct these socio-technical systems. It is not the AI artefact or application that is ethical, trustworthy or responsible. Rather, it is the social component of the socio-technical system that can and should take responsibility and act in consideration of an ethical framework such that the overall system can be trusted by the society. Ethical AI is not, as some may claim, a way to give machines some kind of 'responsibility' for their actions and decisions, and in the process discharge people and organizations of their responsibility. On the contrary, ethical AI gives the people and organizations involved more responsibility and makes them more accountable.

Critique of high-level ethics guidelines

We need to move on from the high-level statements that AI ethics guidelines have produced. They often rely on context-specific keywords — for example, fairness — but do not address the cultural variety between the different societies affected by AI systems[5]. The multi-interpretability of such terms, including that of the concept of AI itself, may prove to be one of the greatest challenges of appropriately regulating intelligent systems[6]. Even if there is a common understanding for some of the terms, to avoid drawn-out semantic debates and minimize risk of adverse outcomes due to misunderstanding, there is a need for disclosure of the contextual definition given to each term. Without a precise framework, developers can easily claim they comply to guidelines and develop ethical AI. At the same time, the opposite may happen; products could be branded unethical (which would destroy consumer confidence) by the company's failure to follow some very loose guidelines. Either outcome could potentially lead to long-term harm of the public's trust.

Actionable policy to assess, develop, incentivize and support the use and

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

forum <privatheit>
selbstbestimmtes_leben_in_der_digitalen_welt

Practical turn

- Practical turn hält an prinzipienbasiertem Zugang fest
- Rahmenwerke sind „lediglich“ detailliertere, feingliedrigere Ethik - Kodizes

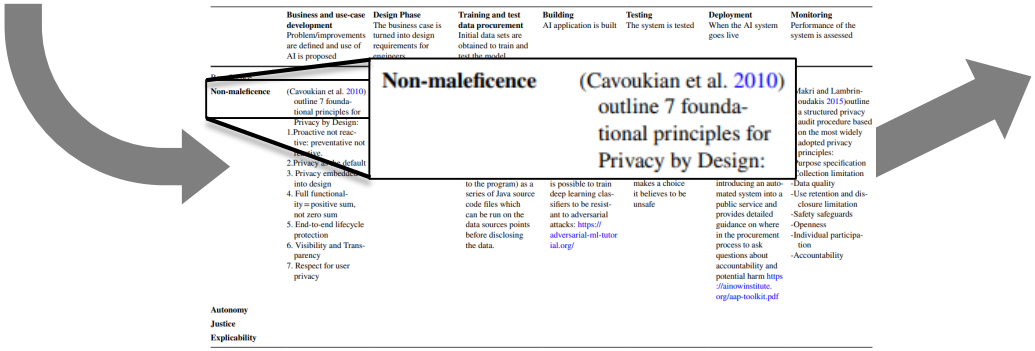
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Beispiel: Privatheit

3- PROTECTION OF PRIVACY AND INTIMACY PRINCIPLE

Privacy and intimacy must be protected from AIS intrusion and data acquisition and archiving systems (DAAS).

source: Montréal Declaration for Responsible Development of Artificial Intelligence



source: Morley et al. (2020) From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices

1. ***Proactive not Reactive; Preventative not Reactive*** Proactive not reactive speaks to the accountability concept of having all the privacy policies as well as mechanisms in place so trained practitioners can observe and resolve privacy issues before they turn into problems.
2. ***Privacy as the Default*** Accountability requires clear organizational rules with an explicit commitment to the policies that are the basis for those rules. Those rules will make clear that information should only be collected and used in a manner that is respectful of individual expectations and a safe information environment.
3. ***Privacy Embedded into Design*** Accountable business processes work best when privacy is embedded into design. This would be part of the mechanisms to implement policies.
4. ***Full Functionality—Positive Sum, Not Zero-Sum*** Organizations that understand privacy and bake privacy in have a better comprehension of the risks to both the organization and to individuals. Organizations that build privacy in know how to create economic value while protecting individual privacy. The Centre purports that clear privacy rules and methodologies create confident organizations that do not suffer from reticence risk.
5. ***End-to-End Lifecycle Protection*** End-to-end lifecycle protection informs the accountable organization that it must build privacy into every process from the assessment before data is collected to the oversight when data is retired.
6. ***Visibility and Transparency*** Principle six requires an organization to be open and honest with individuals. The accountable organization stands ready to demonstrate that it is open about what it practices, stands behind its assertions, and is answerable when questions arise. The accountable organization provides the information necessary for individuals to participate consistent with the OECD individual participation principle. This is echoed in the *Privacy by Design* visibility and transparency principle.
7. ***Respect for User Privacy*** Lastly, the accountable organization must collect, use, store, share and retire information in a manner that is consistent with respect for the individual's privacy.

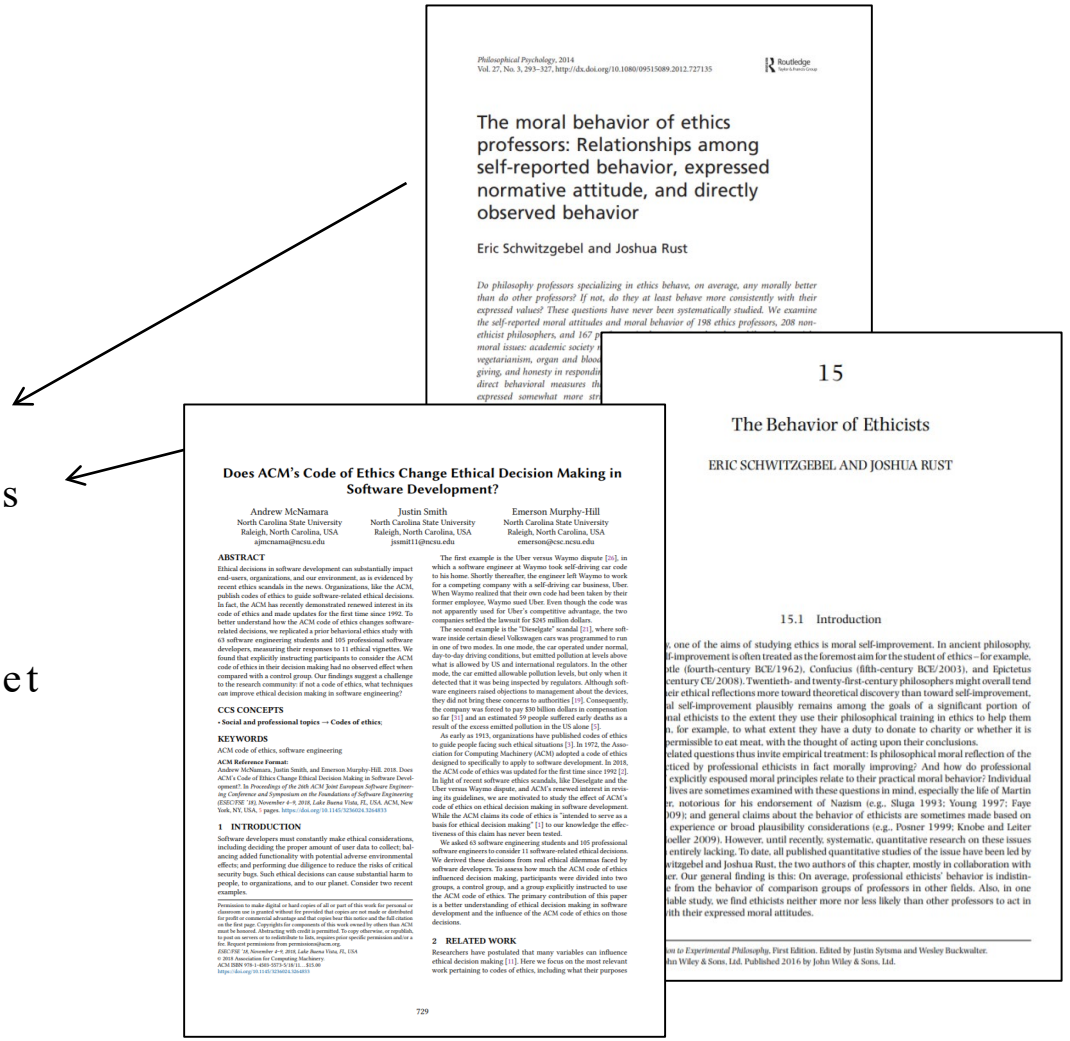
source: Cavoukian et al. (2010) Privacy by Design: essential for organizational accountability and strong business practices

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit> selbstbestimmtes_leben_in_der_digitalen_welt

Unzulänglichkeiten

- practical turn hält an Deontologie fest
- bloße Kenntnis von ethischen Themen hat keinen signifikanten Einfluss auf Entscheidungsfindung
- Der KI-Ethik-Diskurs beachtet psychologische Überlegungen zu bounded ethicality nicht ausreichend



Philosophical Psychology, 2014
Vol. 27, No. 3, 293–327, http://dx.doi.org/10.1080/09515089.2012.727135

The moral behavior of ethics professors: Relationships among self-reported behavior, expressed normative attitude, and directly observed behavior

Eric Schwitzgebel and Joshua Rust

15

The Behavior of Ethicists

ERIC SCHWITZGEBEL AND JOSHUA RUST

15.1 Introduction

Does ACM's Code of Ethics Change Ethical Decision Making in Software Development?

Andrew McNamara
North Carolina State University
Raleigh, North Carolina, USA
ajmcnama@ncsu.edu

Justin Smith
North Carolina State University
Raleigh, North Carolina, USA
jssmit11@ncsu.edu

Emerson Murphy-Hill
North Carolina State University
Raleigh, North Carolina, USA
emerson@csc.ncsu.edu

ABSTRACT

Ethical decisions in software development can substantially impact end-users, organizations, and our environment, as is evidenced by recent ethics scandals in the news. Organizations, like the ACM, publish codes of ethics to guide software-related ethical decisions. In fact, the ACM has recently demonstrated renewed interest in its code of ethics and made updates for the first time since 1992. To better understand how the ACM code of ethics changes software-related decisions, we replicated a prior behavioral ethics study with 63 software engineering students and 105 professional software developers, measuring their responses to 11 ethical vignettes. We found that explicitly instructing participants to consider the ACM code of ethics in their decision making had no observed effect when compared with a control group. Our findings suggest a challenge to the research community: if not a code of ethics, what techniques can improve ethical decision making in software engineering?

CCS CONCEPTS

• Social and professional topics → Codes of ethics;

KEYWORDS

ACM code of ethics, software engineering

ACM Reference Format:
Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. 2018. Does ACM's Code of Ethics Change Ethical Decision Making in Software Development? In Proceedings of the 26th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '18), November 4–9, 2018, Lake Buena Vista, FL, USA. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3236024.3264833

1 INTRODUCTION

Software developers must constantly make ethical considerations, including deciding the proper amount of user data to collect; balancing added functionality with potential adverse environmental effects; and performing due diligence to reduce the risks of critical security bugs. Such ethical decisions can cause substantial harm to people, to organizations, and to our planet. Consider two recent examples.

The first example is the Uber versus Waymo dispute [26], in which a software engineer at Waymo took self-driving car code to his home. Shortly thereafter, the engineer left Waymo to work for a competing company with a self-driving car business, Uber. When Waymo realized that their own code had been taken by their former employee, Waymo sued Uber. Even though the code was not apparently used for Uber's competitive advantage, the two companies settled the lawsuit for $245 million dollars.

The second example is the "Dieselgate" scandal [21], where software inside certain diesel Volkswagen cars was programmed to run in one of two modes. In one mode, the car operated under normal, day-to-day driving conditions, but emitted pollution at levels above what is allowed by US and international regulators. In the other mode, the car emitted allowable pollution levels, but only when it detected that it was being inspected by regulators. Although software engineers raised objections to management about the devices, they did not bring these concerns to authorities [19]. Consequently, the company was forced to pay $30 billion dollars in compensation so far [31] and an estimated 59 people suffered early deaths as a result of the excess emitted pollution in the US alone [5].

As early as 1913, organizations have published codes of ethics to guide people facing such ethical situations [3]. In 1972, the Association for Computing Machinery (ACM) adopted a code of ethics designed to specifically to apply to software development. In 2018, the ACM code of ethics was updated for the first time since 1992 [2]. In light of recent software ethics scandals, like Dieselgate and the Uber versus Waymo dispute, and ACM's renewed interest in revising its guidelines, we are motivated to study the effect of ACM's code of ethics on ethical decision making in software development. While the ACM claims its code of ethics is "intended to serve as a basis for ethical decision making" [2] to our knowledge the effectiveness of this claim has never been tested.

We asked 63 software engineering students and 105 professional software engineers to consider 11 software-related ethical decisions. We derived these decisions from real ethical dilemmas faced by software developers. To assess how much the ACM code of ethics influenced decision making, participants were divided into two groups, a control group, and a group explicitly instructed to use the ACM code of ethics. The primary contribution of this paper is a better understanding of ethical decision making in software development and the influence of the ACM code of ethics on those decisions.

2 RELATED WORK

Researchers have postulated that many variables can influence ethical decision making [11]. Here we focus on the most relevant work pertaining to codes of ethics, including what their purposes

729

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit> selbstbestimmtes_leben_in_der_digitalen_welt

Moralpsychologie in Organisationen

Was sind Antezedens für (un)ethisches Verhalten?

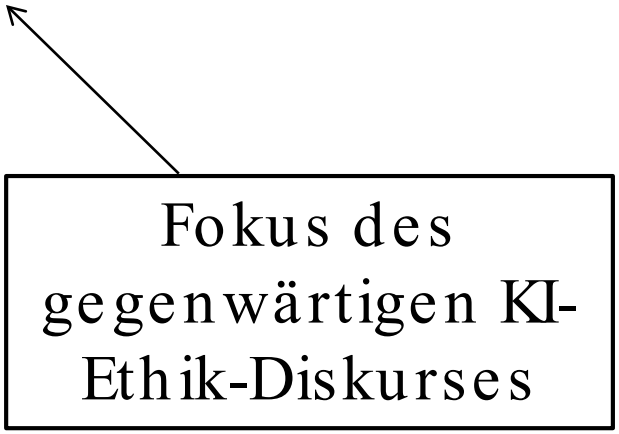
- Merkmale der organisationalen Umwelt
- Moralische Problemstellungen
- Individuelle Eigenschaften

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Moralpsychologie in Organisationen

Was sind Antezedens für (un)ethisches Verhalten?

- **Merkmale der organisationalen Umwelt**
- Moralische Problemstellungen
- Individuelle Eigenschaften

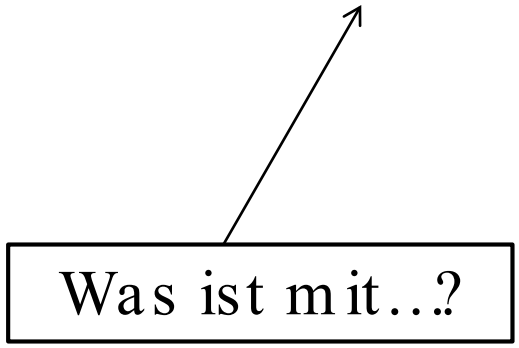


GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Moralpsychologie in Organisationen

Was sind Antezedens für (un)ethisches Verhalten?

- Merkmale der organisationalen Umwelt
- Moralische Problemstellungen
- **Individuelle Eigenschaften**



GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

forum
<privatheit>
selbstbestimmtes_leben_
in_der_digitalen_welt

Übergang

(externe) Prinzipien

- Deontologie
- universelle normative Regeln
- moralische Pflichten
- handlungsleitende Prinzipien

(interne) Tugenden

- Tugendethik
- Motivationen
- Charakterzüge/Disposition
- Persönlichkeitsentwicklung

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Übergang

(externe) Prinzipien

- Deontologie
- Uniuniverselle normative Regeln
- Moralische Pflichten
- Handlungsleitende Prinzipien

(interne) Tugenden

- Tugendethik
- Motivationen
- Charakterzüge/Dispositionen
- Persönlichkeitsentwicklung

technomoralische Tugenden
(Vallor 2016)

- Mut
- Bescheidenheit
- Großzügigkeit
- usw.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

KI-spezifische Tugenden

Wie können KI-spezifische Tugenden abgeleitet werden?

- Nutzung von Meta-Analysen über Richtlinien zur KI-Ethik
- Prinzipien mit Tugenden korrelieren

AI virtues

The missing link in putting AI ethics into practice

Dr. Thilo Hagendorff
thilo.hagendorff@uni-tuebingen.de
University of Tuebingen
Cluster of Excellence "Machine Learning: New Perspectives for Science"
International Center for Ethics in the Sciences and Humanities

Abstract – Several seminal ethics initiatives have stipulated sets of principles and standards for good technology development in the AI sector. However, widespread criticism has pointed out a lack of practical realization of these principles. Following that, AI ethics underwent a practical turn, but without deviating from the principled approach and the many shortcomings associated with it. This paper proposes a different approach. It defines four basic AI virtues, namely justice, honesty, responsibility and care, all of which represent specific motivational settings that constitute the very precondition for ethical decision making in the AI field. Moreover, it defines two second-order AI virtues, prudence and fortitude, that bolster achieving the basic virtues by helping with overcoming bounded ethicality or the many hidden psychological forces that impair ethical decision making and that are hitherto disregarded in AI ethics. Lastly, the paper describes measures for successfully cultivating the mentioned virtues in organizations dealing with AI research and development.

Keywords – AI virtues; AI ethics; business ethics; moral psychology; bounded ethicality; implementation; machine learning; artificial intelligence

1 Introduction

Current AI ethics initiatives, especially when adopted in scientific institutes or companies, embrace a principle-based, deontological approach (Mittelstadt 2019). However, establishing principles alone does not suffice, they also must be convincingly put into practice. Most AI ethics guidelines do shy away from coming up with methods to accomplish this (Hagendorff 2020c). Nevertheless, recently more and more research papers appeared that describe steps on how to come "from what to how" (Morley et al. 2020; Vakkuri et al. 2019a; Eitel-Porter 2020; Theodorou and Dignum 2020). However, AI ethics still fails in certain regards. The reasons for that are manifold and reach from economical or legal to various other socio-cultural constraints. Economic imperatives in particular can overwrite ethical concerns and intentions. This is why both in academia and public debates, many authors state that AI ethics has not permeated the AI industry, quite the

1

Welche Tugend *A*, *B*, *C* schlägt sich in einem Verhalten nieder, das mit hoher Wahrscheinlichkeit zu einem Resultat führt, welches den Erfordernissen des Prinzips *X*, *Y*, *Z* entspricht?

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit>
selbstbestimmtes_leben_in_der_digitalen_welt

Meta-Analysen

Minds and Machines (2020) 30:99–120
https://doi.org/10.1007/s11023-020-09517-8

Check for updates

The Ethics of AI Ethics: An Evaluation of Guidelines

Thilo Hagendorff[1]

Received: 1 October 2019 / Accepted: 21 January 2020 / Published online: 1 February 2020

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 semi-systematic evaluation, this paper analyzes and compares 22 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 22 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

✉ Thilo Hagendorff
thilo.hagendorff@uni-tuebingen.de

[1] Cluster of Excellence "Machine Learning: New Perspectives for Science", University of Tuebingen, Tübingen, Germany

Springer

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Research Publication No. 2020-1
January 15, 2020

Principled Artificial Intelligence:
Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI

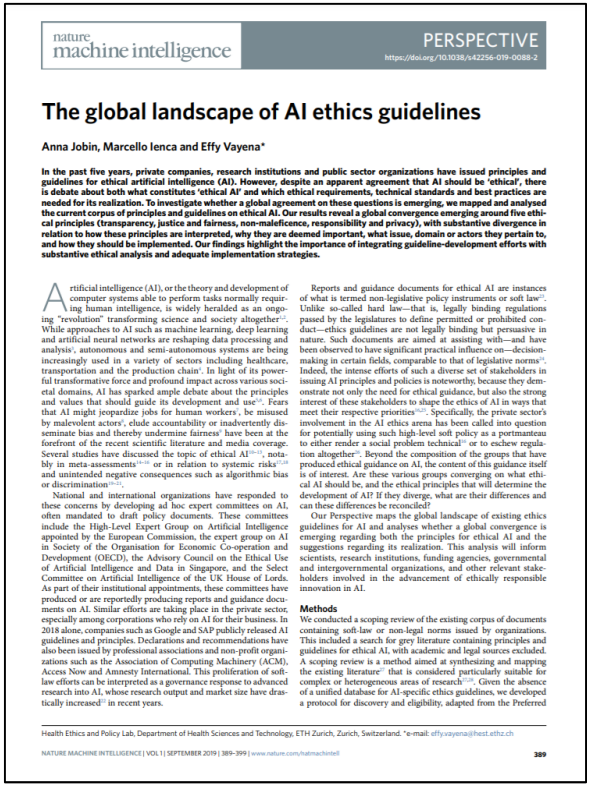
Jessica Fjeld
Nele Achten
Hannah Hilligoss
Adam Christopher Nagy
Madhulika Srikumar

This paper can be downloaded without charge at:

The Berkman Klein Center for Internet & Society Research Publication Series:
https://cyber.harvard.edu/publication/2020/principled-ai

The Social Science Research Network Electronic Paper Collection:
https://ssrn.com/abstract=3518482

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax) • http://cyber.law.harvard.edu/ • cyber@law.harvard.edu



nature machine intelligence
PERSPECTIVE
https://doi.org/10.1038/s42256-019-0088-2

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca and Effy Vayena*

In the past five years, private companies, research institutions and public sector organizations have issued principles and guidelines for ethical artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. To investigate whether a global agreement on these questions is emerging, we mapped and analysed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented. Our findings highlight the importance of integrating guideline-development efforts with substantive ethical analysis and adequate implementation strategies.

Artificial intelligence (AI), or the theory and development of computer systems able to perform tasks normally requiring human intelligence, is widely heralded as an ongoing "revolution" transforming science and society altogether[1,2]. While approaches to AI such as machine learning, deep learning and artificial neural networks are reshaping data processing and analysis[3], autonomous and semi-autonomous systems are being increasingly used in a variety of sectors including healthcare, transportation and the production chain[4]. In light of its powerful transformative force and profound impact across various societal domains, AI has sparked ample debate about the principles and values that should guide its development and use[5,6]. Fears that AI might jeopardize jobs for human workers[7], be misused by malevolent actors[8], elude accountability or inadvertently disseminate bias and thereby undermine fairness[9] have been at the forefront of the recent scientific literature and media coverage. Several studies have discussed the topic of ethical AI[10–13], notably in meta-assessments[14–17] or in relation to systemic risks[18–20] and unintended negative consequences such as algorithmic bias or discrimination[10–21].

National and international organizations have responded to these concerns by developing ad hoc expert committees on AI, often mandated to draft policy documents. These committees include the High-Level Expert Group on Artificial Intelligence appointed by the European Commission, the expert group on AI in Society of the Organisation for Economic Co-operation and Development (OECD), the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore, and the Select Committee on Artificial Intelligence of the UK House of Lords. As part of their institutional appointments, these committees have produced or are reportedly producing reports and guidance documents on AI. Similar efforts are taking place in the private sector, especially among corporations who rely on AI for their business. In 2018 alone, companies such as Google and SAP publicly released AI guidelines and principles. Declarations and recommendations have also been issued by professional associations and non-profit organizations such as the Association of Computing Machinery (ACM), Access Now and Amnesty International. This proliferation of soft-law efforts can be interpreted as a governance response to advanced research into AI, whose research output and market size have drastically increased[22] in recent years.

Reports and guidance documents for ethical AI are instances of what is termed non-legislative policy instruments or soft law[23]. Unlike so-called hard law—that is, legally binding regulations passed by the legislatures to define permitted or prohibited conduct—ethics guidelines are not legally binding but persuasive in nature. Such documents are aimed at assisting with—and have been observed to have significant practical influence on—decision-making in certain fields, comparable to that of legislative norms[24]. Indeed, the intense efforts of such a diverse set of stakeholders in issuing AI principles and policies is noteworthy, because they demonstrate not only the need for ethical guidance, but also the strong interest of these stakeholders to shape the ethics of AI in ways that meet their respective priorities[16,25]. Specifically, the private sector's involvement in the AI ethics arena has been called into question for potentially using such high-level soft policy as a portmanteau to either render a social problem technical[26] or to eschew regulation altogether[27]. Beyond the composition of the groups that have produced ethical guidance on AI, the content of this guidance itself is of interest. Are these various groups converging on what ethical AI should be, and the ethical principles that will determine the development of AI? If they diverge, what are their differences and can these differences be reconciled?

Our Perspective maps the global landscape of existing ethics guidelines for AI and analyses whether a global convergence is emerging regarding both the principles for ethical AI and the suggestions regarding its realization. This analysis will inform scientists, research institutions, funding agencies, governmental and intergovernmental organizations, and other relevant stakeholders involved in the advancement of ethically responsible innovation in AI.

Methods

We conducted a scoping review of the existing corpus of documents containing soft-law or non-legal norms issued by organizations. This included a search for grey literature containing principles and guidelines for ethical AI, with academic and legal sources excluded. A scoping review is a method aimed at synthesizing and mapping the existing literature[28] that is considered particularly suitable for complex or heterogeneous areas of research[29,30]. Given the absence of a unified database for AI-specific ethics guidelines, we developed a protocol for discovery and eligibility, adapted from the Preferred

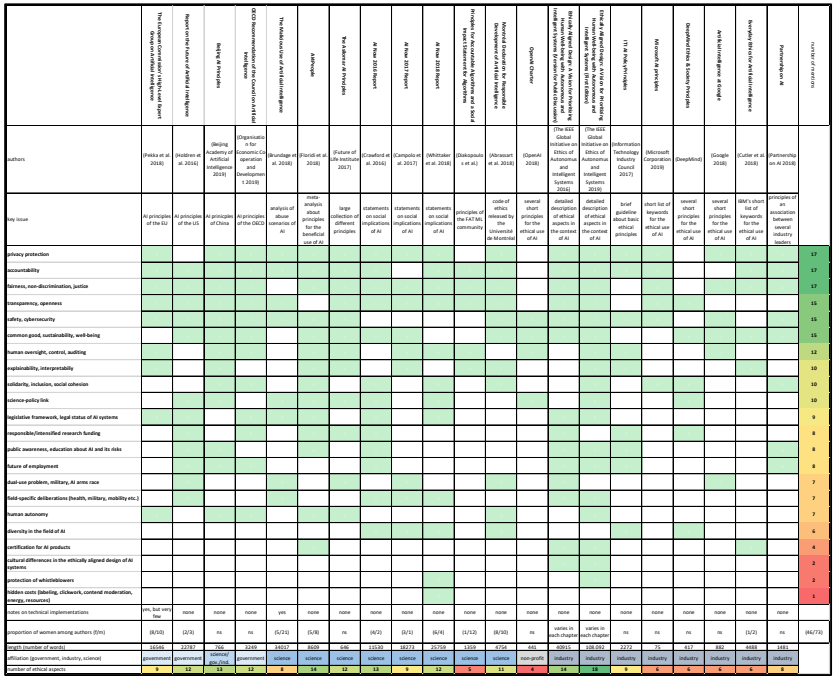
Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. *e-mail: effy.vayena@hest.ethz.ch

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

forum
<privatheit>
selbstbestimmtes_leben_
in_der_digitalen_welt

Studie 1

Hagendorff, Thilo (2020): The Ethics of AI Ethics. An Evaluation of Guidelines. In Minds and Machines 30 (3), pp. 457–461.



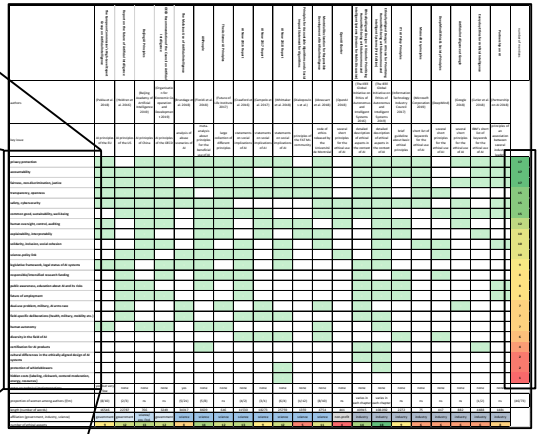
	The European Commission's High-level Expert Group on Artificial Intelligence	Report on the Future of Artificial Intelligence	Beijing Academy of Artificial Intelligence	(Organisation for Economic Co-operation and Development)	The Public Voice on Artificial Intelligence	ASILOMAR	The Montreal Declaration for a Responsible Development of Artificial Intelligence	AI Now 2016 Report	AI Now 2017 Report	AI Now 2018 Report	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Asilomar Principles of Artificial Intelligence	Ethically Aligned Design	IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	Ethically Aligned Design, A Vision For Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition	ITI AI Policy Principles	Microsoft AI principles	DeepMind Ethics & Society Principles	Ad Hoc Committee on AI (CAHAI)	Google AI Principles	Cutler et al. Harvard Business Review	Partnership on AI	number of criteria
authors	(Pekka et al. 2018)	(Haldren et al. 2016)	(Beijing Academy of Artificial Intelligence 2019)	(Organisation for Economic Co-operation and Development 2019)	(Brundage et al. 2018)	(Floridi et al. 2018)	(Future of Life Institute 2017)	(Crawford et al. 2016)	(Campolo et al. 2017)	(Whittaker et al. 2018)	(Diakopoulos et al.)	(Abrassart et al. 2018)	(OpenAI 2018)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2016)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019)	(Information Technology Industry Council 2017)	(Microsoft Corporation 2018)	(DeepMind)	(Google 2018)	(Cutler et al. 2018)	(Partnership on AI 2018)	
key issue	AI principles of the EU	AI principles of the US	AI principles of China	AI principles of the OECD	analysis of abuse scenarios of AI	meta-analysis about principles for the beneficial use of AI	large collection of different principles	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	principles of the FAT ML community	code of ethics released by the Université de Montréal	several short principles for the ethical use of AI	detailed description of ethical aspects in the context of AI	detailed description of ethical aspects in the context of AI	brief guideline about basic ethical principles	short list of keywords for the ethical use of AI	several short principles for the ethical use of AI	several short principles for the ethical use of AI	BMW's short list of keywords for the ethical use of AI	principles of an association between several industry leaders	
privacy protection																							17
accountability																							17
fairness, non-discrimination, justice																							17
transparency, openness																							15
safety, cybersecurity																							15
common good, sustainability, well-being																							15
human oversight, control, auditing																							12
explainability, interpretably																							10
solidarity, inclusion, social cohesion																							10
science-policy link																							10
legislative framework, legal status of AI systems																							9
responsible/intensified research funding																							8
public awareness, education about AI and its risks																							8
future of employment																							8
dual-use problem, military, AI arms race																							7
field-specific deliberations (health, military, mobility etc.)																							7
human autonomy																							7
diversity in the field of AI																							6
certification for AI products																							4
cultural differences in the ethically aligned design of AI systems																							3
protection of whistleblowers																							2
hidden costs (labeling, clickwork, contend moderation, energy, resources)																							1
notes on technical implementations	yes, but very few	none	none	none	yes	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none		
proportion of women among authors (f/m)	(8/50)	(2/3)	ns	ns	(1/21)	(1/8)	ns	(4/2)	(3/1)	(6/4)	(1/12)	(8/30)	ns	varies in each chapter	varies in each chapter	ns	ns	ns	ns	(1/2)	ns	(44/73)	
length (number of words)	16546	22787	766	2249	34927	8609	646	11530	18273	25759	1359	4754	449	40915	109.092	2272	75	417	882	4488	1483		
affiliation (government, industry, science)	government	government	science and Govt	government	science	science	science	science	science	science	science	science	non-profit	industry	industry	industry	industry	industry	industry	industry	industry		
number of ethical aspects	9	12	13	12	8	14	12	13	9	12	5	11	8	14	18	9	6	6	6	6	8		

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit> selbstbestimmtes_leben_in_der_digitalen_welt

Studie 1

privacy protection	17
accountability	17
fairness, non-discrimination, justice	17
transparency, openness	15
safety, cybersecurity	15
common good, sustainability, well-being	15
human oversight, control, auditing	12
explainability, interpretabiliy	10
solidarity, inclusion, social cohesion	10
science-policy link	10
legislative framework, legal status of AI systems	9
responsible/intensified research funding	8
public awareness, education about AI and its risks	8
future of employment	8
dual-use problem, military, AI arms race	7
field-specific deliberations (health, military, mobility etc.)	7
human autonomy	7
diversity in the field of AI	6
certification for AI products	4
cultural differences in the ethically aligned design of AI systems	2
protection of whistleblowers	2
hidden costs (labeling, clickwork, contend moderation, energy, resources)	1

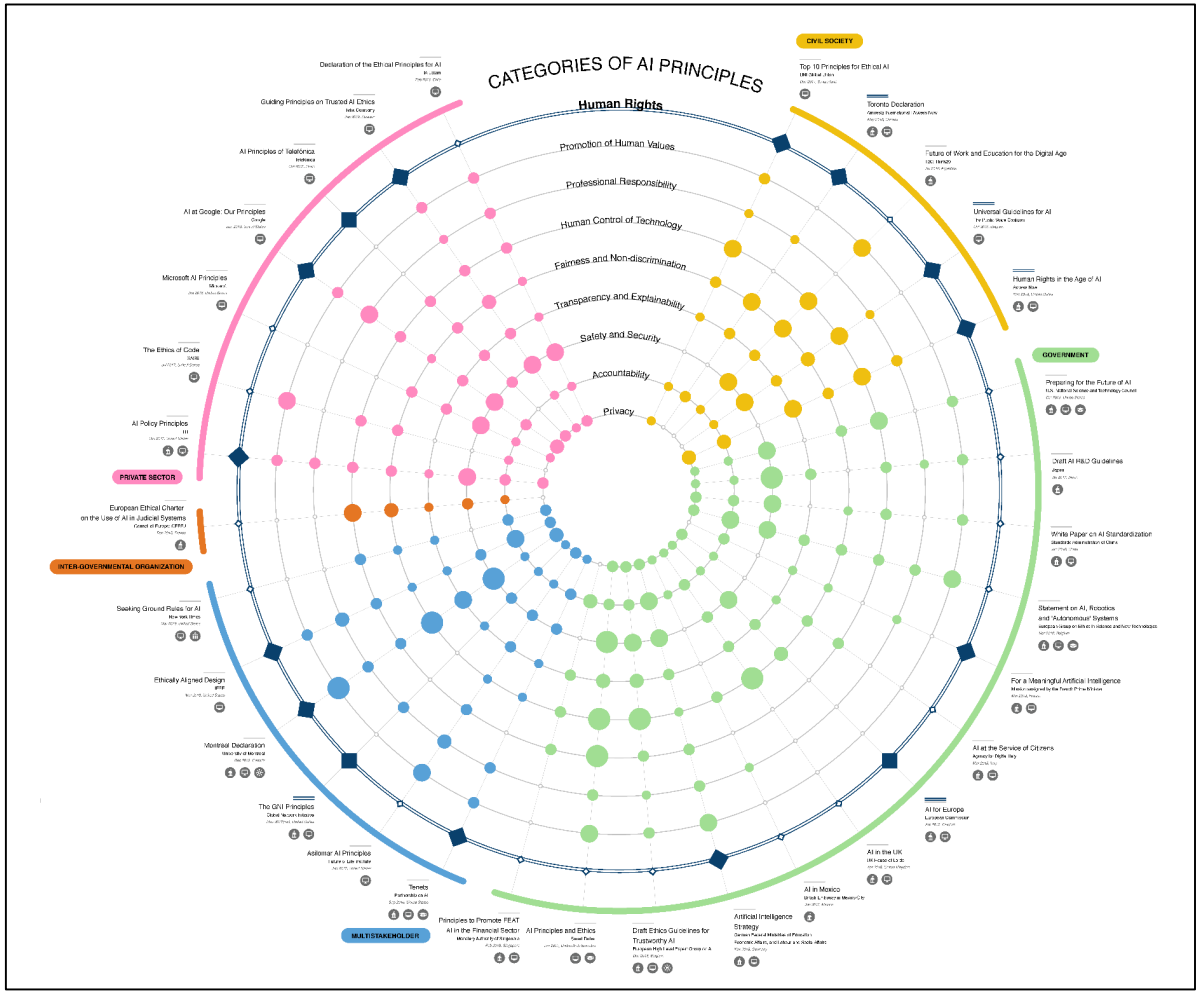


GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit> selbstbestimmtes_leben_in_der_digitalen_welt

Studie 2

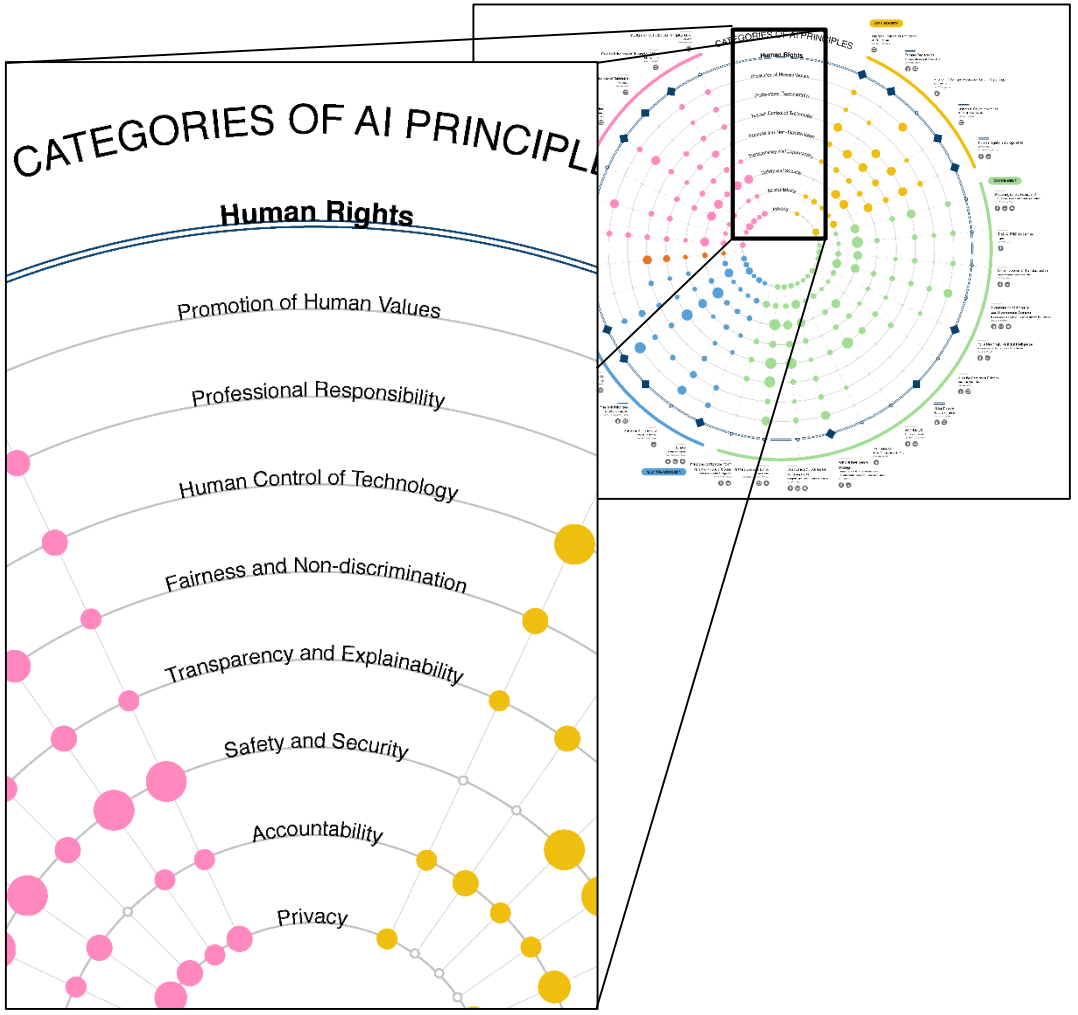
Fjeld, Jessica; Achten, Nele; Hilligoss, Hannah; Nagy, Adam; Srikumar, Madhulika (2020): Principled Artificial Intelligence. Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research Publication No. 2020-1. In SSRN Journal, pp. 1–39.



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit> selbstbestimmtes_leben_in_der_digitalen_welt

Studie 2



GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

forum <privatheit>
selbstbestimmtes leben in der digitalen welt

Studie 3

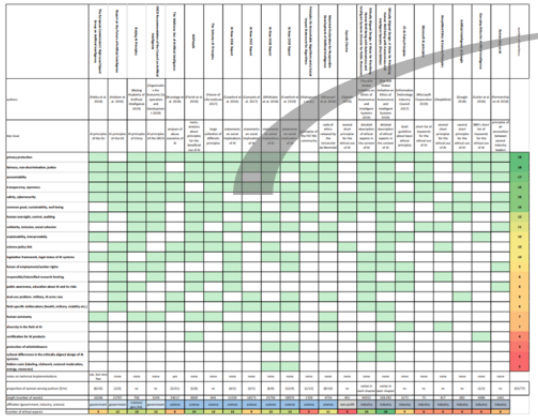
Jobin, Anna; Ienca, Marcello; Vayena, Effy (2019): The global landscape of AI ethics guidelines. In Nature Machine Intelligence 1 (9), pp. 389–399.

Table 3 | Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Tugenden



Hagendorff 2020c

Jobin et al. 2019

Table 3 | Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Fjeld et al. 2020

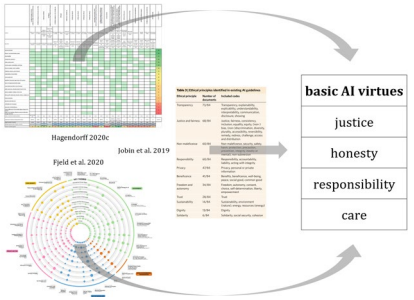
basic AI virtues
justice
honesty
responsibility
care

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit> selbstbestimmtes_leben_in_der_digitalen_welt

Basale KI-Tugenden

Prinzipien	Entsprechende Tugenden
Algorithmische Fairness, Nichtdiskriminierung, Abschwächung von Biases, Inklusion, Gleichheit, diversity/Vielfalt	**Gerechtigkeit (justice)**
Organisationale Transparenz, Offenheit, Erklärbarkeit, Interpretierbarkeit, open source, Anerkennung von Fehlern und Irrtümern	**Ehrlichkeit (honesty)**
Verantwortung, Verbindlichkeit, Haftbarkeit, Replizierbarkiet, Legalität, Genauigkeit, Berücksichtigung (langfristiger) technologischer Konsequenzen	**Verantwortung (responsibility)**
Unbedenklichkeit, Schaden, Sicherheit, Privatheit, Schutz, Vorsorge, versteckte Kosten, Wohlergehen, Nachhaltigkeit, Frieden, Gemeinwohl, Solidarität, soziale Kohäsion, Freiheit, Autonomie, Einwilligung	**Sorge (care)**



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit>
selbstbestimmtes_leben_in_der_digitalen_welt

Kritische Fragen

- Warum ist dieser Zugang besser als die alten?
- Ist dieser nicht noch viel abstrakter und weniger praxisbezogen?
- Wo sind die technischen Details?

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

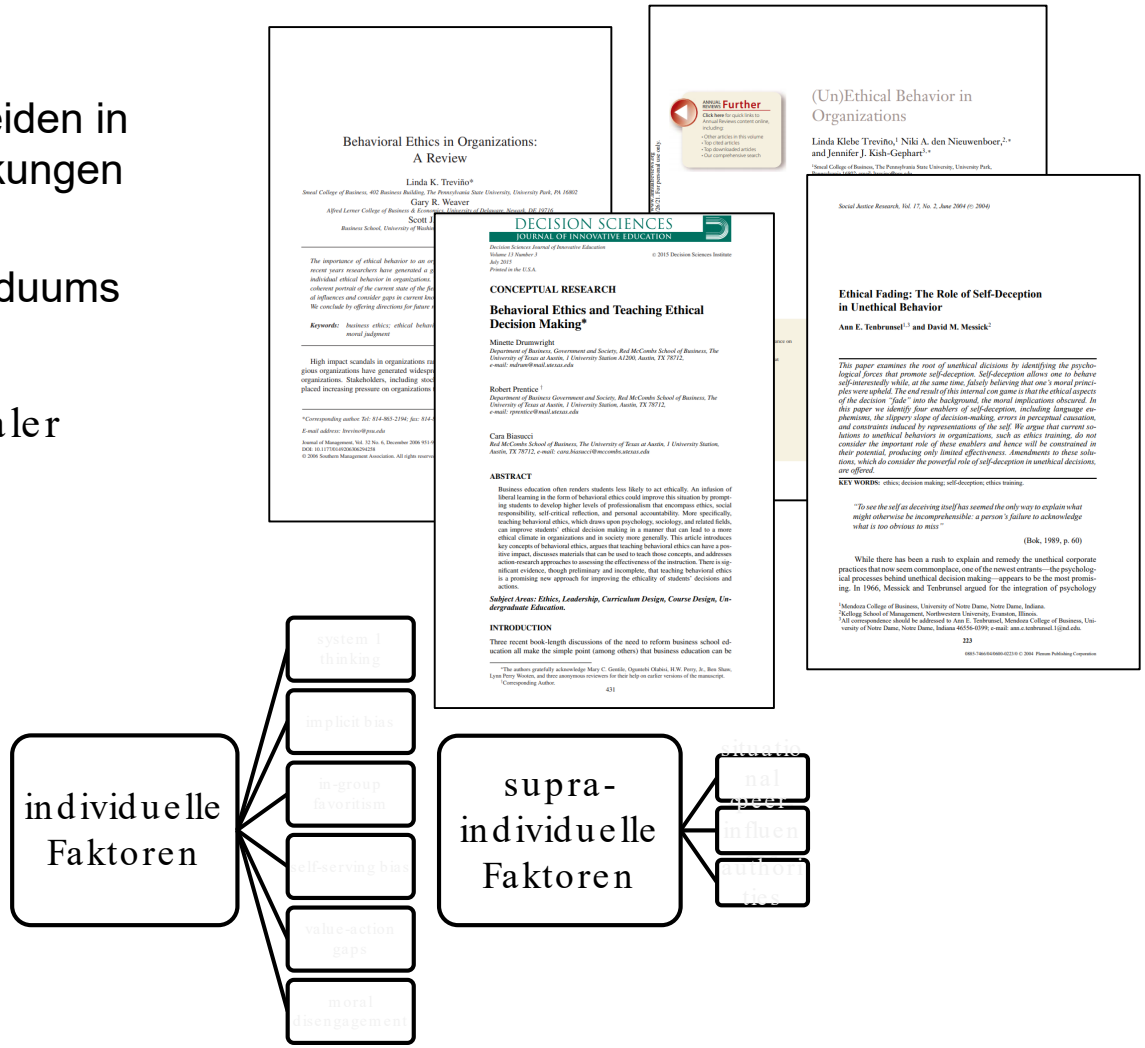
War es das?

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

forum
<privatheit>
selbstbestimmtes_leben_
in_der_digitalen_welt

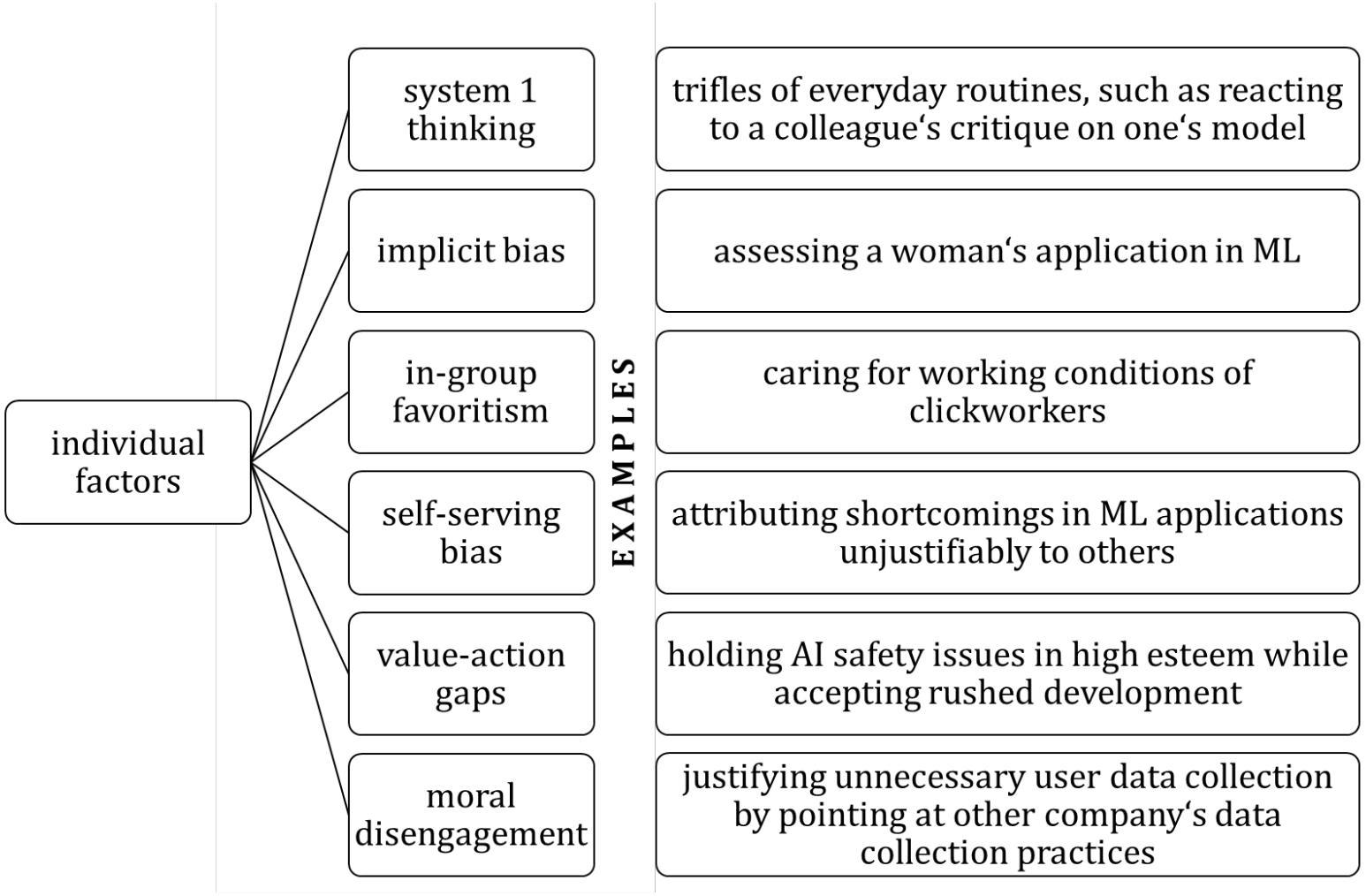
War es das?

- Trotz existierender basaler KI - Tugenden ist ethisches Entscheiden in der Praxis mit vielen Einschränkungen konfrontiert
- „bounded ethicality “ des Individuums
  - kognitive Fehler, Biases
  - sozialer oder organisationaler Druck
  - situative Kräfte
  - etc.

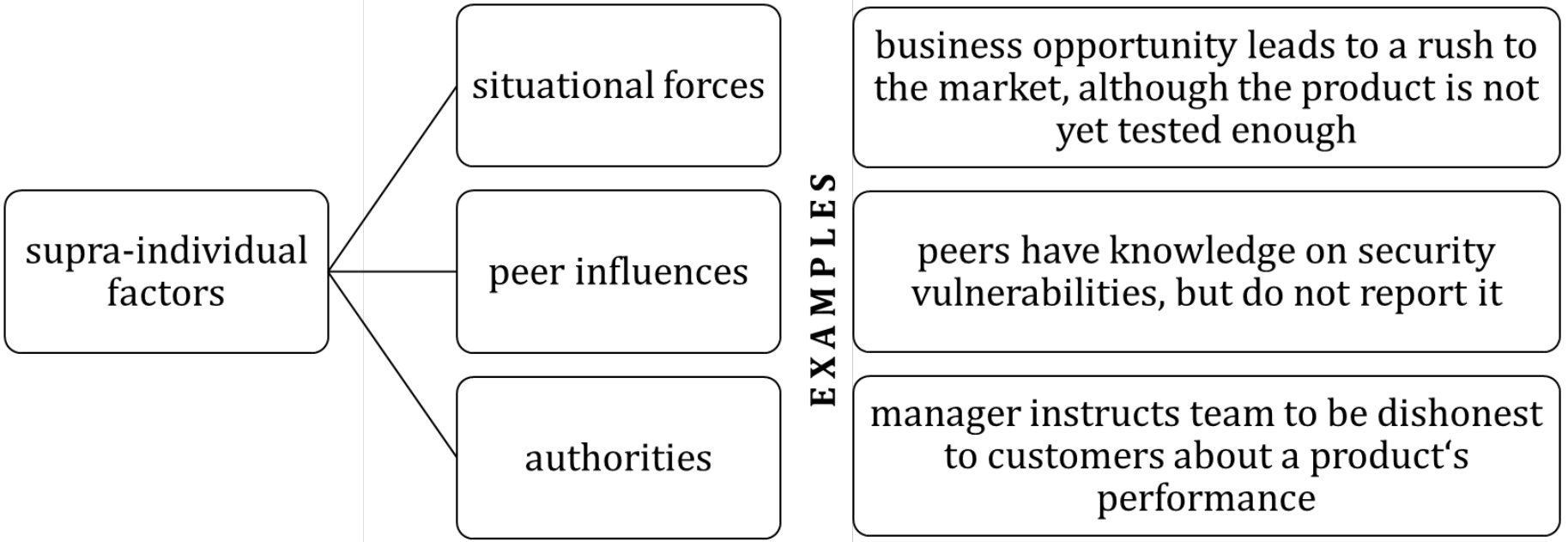


GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

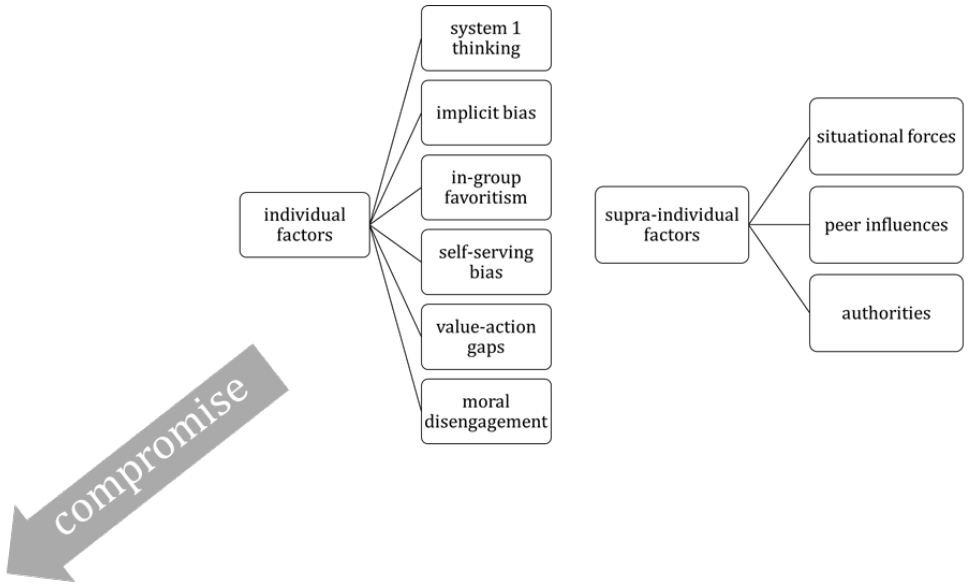
forum <privatheit> selbstbestimmtes_leben_in_der_digitalen_welt



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

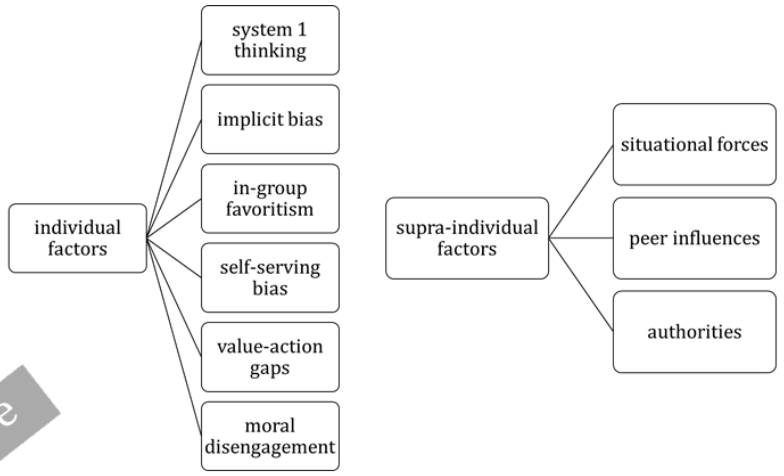


GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung



basic AI virtues
justice
honesty
responsibility
care

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung



compromise

basic AI virtues
justice
honesty
responsibility
care

addition

second-order AI virtues
prudence
fortitude

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

forum <privatheit>
selbstbestimmtes_leben_in_der_digitalen_welt

KI-Tugenden zweiter Ordnung

bounded ethicality	Tugenden
system 1 thinking, implizite Voreingenommenheit/Bias, in-group-Favorisierung, self-serving bias, value-action gaps, moralische Distanzierung, usw.	**Prudence**
Situative Kräfte, Peer-Beeinflussung Autoritäten, usw.	**Fortitude**

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

KI-Tugenden

Basale KI -Tugenden	KI-Tugenden zweiter Ordnung	Tugenden kultivieren
justice honesty responsibility care	prudence fortitude	?

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

forum
<privatheit>
selbstbestimmtes_leben_
in_der_digitalen_welt

Ethik-Training

- KI-Tugenden in einem spezifischen organisationalen Kontext etablieren und kultivieren
- Tugenden können trainiert und gefördert werden

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Maßnahmen

Individuelle Ebene

- Wissen über KI-Tugenden
- Handlungsstrategien
- Locus of control
- öffentliche Bekenntnis
- Audits und Diskussionsrunden

Systemische Ebene

- Führungskräfte
- Ethische Organisationskultur
- Frauenquote
- Stress und Druck reduzieren
- Offenheit für Kritik

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

forum
<privatheit>
selbstbestimmtes_leben_
in_der_digitalen_welt

Kritische Stimmen

- Forderungen nach praktischer Umsetzung von KI-Ethik
- Wie kann dies erreicht werden?

6 Reasons Why AI Ethics in Corporations is All Talk and No Action

source: https://becominghuman.ai/6-reasons-why-ai-ethics-in-corporations-is-all-talk-and-no-action-4f126af42668

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Mein Vorschlag

- Prinzipienlisten überwinden
- Moralpsychologie berücksichtigen

Basale KI-Tugenden
+
KI-Tugenden zweiter Ordnung
+
Ethik-Training

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Danke!

Dr. Thilo Hagendorff

thilo.hagendorff@uni -tuebingen.de

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

forum <privatheit>
selbstbestimmtes_leben_
in_der_digitalen_welt