VISIONS OF HUMAN-CENTERED ARTIFICIAL INTELLIGENCE – RELATIONS WITH ETHICS AND POWER

Lilla Vicsek and Tamás Tóth

Preproof version of the chapter: Vicsek, L., & Tóth, T. (2023). Visions of Human-Centered Artificial Intelligence: Relations with Ethics and Power. In *Algorithmic Ethics* (pp. 1-21). Routledge.

ABSTRACT

In recent years there has been heightened sensibility to ethical issues connected with artificial intelligence. Conceptions of different types of AI have emerged which touch upon this new sensibility, including wholesome augmented intelligence, responsible/trustworthy/ethical AI, AI for good, and human-centered AI. In this chapter, we focus on the concept of human-centered AI, which has gained more prominence lately and which has even appeared in the names of organizations. However, there are major problems with the conceptualization and operationalization of human-centered AI. This chapter critically analyzes academic visions about human-centered AI in five Western university institutions' online textual content (n=573). The study scrutinizes institutions that use the term "human-centered" in their names. Even though institutions provide more content framed with supportive attitudes rather than focusing on technical solutions, the related texts fail to address several important issues. First, they often treat humanity as a homogenous group, suggesting that every society struggles with the same problems. Second, human-centered AI is treated as being mainly aligned with the Global North's needs. Finally, most of the texts associated with the scrutinized institutions lack discussion of the surging inequalities connected to the capitalist system. Therefore, they do not offer many AI-supported solutions that might address the challenges of a lack of clean water, poverty, or the presence of dangerous jobs that harm the Global South. Instead, the analyzed institutions primarily present societal challenges within national borders, while they disregard the need for redressing fundamental problems that hinder the creation of acceptable living conditions in poor regions. Building on the sociology of expectations, this study argues that the visions of human-centered AI are of paramount importance. These expectations have the potential to legitimize, guide, and coordinate the activities of different actors responsible for the research, development, and application of AI-driven technologies.

Keywords: AI ethics, human-centered artificial intelligence, human in the loop, Global North, Global South

INTRODUCTION

There has been heightened attention to the disruptive potentials of AI solutions and ethical issues lately. Leading companies in AI development have also begun to deal with ethical problems. However, the way they and other organizations have treated these has received severe criticisms from social scientists. It seems ethics statements often serve the purpose of calming critical voices and reassuring the public, investors, and legislators, whilst deflecting regulation of the sphere (Kerr et al. 2020). Additionally, ethical discourses often cannot influence AI development and decision-making to a high degree as ethical considerations tend to be too general and not concrete enough for effective application (Kerr et al. 2020, Hagendorff 2020).

Greene et al. (2019) have pointed out that high-profile ethical statements of major independent organizations between 2015 and 2018 did not question the “status quo”: the current social and business arrangements under which AI was developed. This was typical of the statements, despite the fact that mainly huge companies without proper “democratic oversight” and with underpaid, insecure workforce develop these technologies (Greene et al 2019). Other activists and social scientists (Dotson 2015) have claimed that the development of a profit-oriented technology did not always lead to the best solutions for humanity and could be harmful, especially if it was not regulated strongly enough.

Several authors have criticized the principles and values in ethical statements on AI as often representing Western neoliberal principles only and that in applying them they can contribute to global inequality (Stark et al. 2021, Monahan 2021). Monahan (2021) has criticized the transparency ideal, especially in the case of surveillance, as contributing to the solidifying of Western white male supremacy. He argues, based on Haraway (1991), that science has often been tied to militarism, capitalism, colonialism, male supremacy. Much research in Science and Technology Studies (STS) has demonstrated how technological solutions often serve certain social goals of a specific privileged groups. Authors in STS often emphasize there are choices regarding technological directions and the resulting solutions can benefit some groups, whilst not benefiting others. Moreover, the funds allocated to technological development are taking away money from other ways to help social problems.

Hagendorff (2020) has found in his analysis of 22 ethical guidelines that frequently mentioned aspects – such as accountability, explainability, privacy, justice, robustness, safety were often executed as technical solutions. Based on Gilligan (1982), he asserts these technical solutions can be regarded as instances of male-dominated calculating and logic-oriented ethics, and the guidelines leave out non-masculine ways of thinking, including discussing AI ethics from the aspects of care, nurture, help, welfare, social responsibility, etc. “In AI ethics, technical artifacts are primarily seen as isolated entities that can be optimized by experts so as to find technical solutions for technical problems” (Hagendorff 2020, p. 112). He found that even in the field of AI ethics most authors are men, and that the problem of lack of diversity of AI developers is missing from many guidelines.

One term that has appeared to defend AI against criticisms is “human-centered AI”. Projects and research centers use the term in their names. Additionally, university courses, MSc programs, companies, blogs refer to human-centered AI and EU Horizons mentions it as a priority. Therefore, we find it relevant to analyze visions of human-centered AI. We aim to examine visions of human-centered artificial intelligence (HAI); what these visions leave out and what they include, responding sensitively to issues of power and inequality within and between societies. We compare these visions with features of previous ethical value statements and guidelines. We also discuss the consequences of these expectations.

Our study is built on the following premises of the sociology of technological expectations: (1) anticipations to technologies are important aspects of modern capitalism (Beckert 2016), (2) visions of technologies have a constitutive role as they can influence action, legitimize, show direction and coordinate actions of actors (Van Lente 2012), (3) powerful actors can influence future expectations in science and technologies and thus shape the future by trying to marginalize alternative channels of future development (Brown et al., 2000). In this light, we investigate the HAI imaginaries of specific agents, namely university "institutes" to introduce their understanding of HAI concerning national and supranational levels.[1]

METHODOLOGY

Our goal was to analyze five university institutions' visions of HAI from the Global North (Demeter, 2020); therefore, we conducted qualitative, thematic content analysis on texts published via their official websites. We scrutinized five institutes (see Table 1) from two European and three American universities with "Human-centered" and "AI" terms in their names.

Our database consists of the institutions' introductions, goals, visions, staff bios, published events, lectures, seminars, news, reports, interviews, calls for papers, and grants. Note that as extended reports published via Stanford University's website were extraordinarily long, we scrutinized only their introductions and conclusions. The total number of analyzed documents is 573. We started the data collection on 6 July 2021 and finished it on 6 October 2021, starting with the first published content from 2018.

Table 1 – Analyzed university' HAI institutes or research areas

University	Institute	Hosting Country	Text no. & share of the sample
University of Bologna	Alma Mater Research Institute for Human-Centered Artificial Intelligence	Italy	n=32 (5.6%)
Utrecht University	Human-centered Artificial Intelligence focus area	Netherlands	n=59 (10.3%)
Stanford University	Institute for Human-Centered AI	United States	n=352 (61.4%)
Northwestern University	Center for Human-Computer Interaction + Design, Human-Centered AI research area	United States	n=36 (6.3%)
University of Maryland	Human-Computer Interaction Laboratory, Human-Centered AI research area	United States	n=94 (16.4%)

In our scrutiny, we operationalized the following definition for a theme: it "captures something important about the data in relation to the research question, and represents some level of patterned response or meaning within the data set" (Braun & Clarke, 2006, p. 82). We analyzed the database using theoretical thematic analysis that primarily relied on Hagendorff's (2020) categorization of AI ethics. "Supportive attitudes," "Power issues," and "Technical solutions" are three themes that are intertwined with Hagendorff's work, while "Vulnerable Groups" and "Capitalism" rely on other research on AI and economy (Fountain, 2021; Piketty, 2020; Tiyasha et al., 2020; Xiang et al., 2021). We provided all themes and subthemes utilized during the analysis (see Table 2), which focused on the interpretative level to understand the themes' universal connotations and implications concerning contemporary research on AI ethics (Hagendorff, 2020; Patton, 1990). We note that "Vulnerable groups" should fall under "Supportive attitudes" but methods-wise, our separation relies on the features of "Supportive attitudes" based on former ethical guidelines that mostly lack marginalized circles (Hagendorff, 2020). Therefore, we created a unique category for vulnerable entities.

We chose paragraphs as coding units to avoid thematic discontinuities (Rooduijn, 2014). We coded every theme within the specific paragraph where topics were perceived. The themes'

[1] For the sake of coherence and simplicity, we refer to departments, institutions, research labs, university units, research areas, etcetera utilizing "Human-centered" and "AI" in their names as "institutes" or "institutions."

structure is the following: we created five themes and thirteen subthemes. We marked themes with numbers ranging between 1-5 and utilized letters with small captions to introduce subthemes (see Table 2-4). The rest of this section briefly introduces our categories to characterize the typology that helped us conduct the thematic content analysis.

Table 2 – Structure and frequencies of themes and subthemes

Themes	**Frequency**
1) Capitalism	*16*
2) Vulnerable groups	*420*
3) Power issues (decision-making)	*673*
4) Supportive attitudes	*1,704*
a) Social responsibility & sustainability	206
b) Welfare	180
c) Help	488
d) Care & nurture	830
5) Technical solutions	*1,049*
e) Safety	223
f) Robustness	44
g) Explainability	29
h) Accountability	48
i) Privacy	161
j) Responsibility	95
k) Non-maleficence (causing no harm)	48
l) Justice & fairness	344
m) Transparency	57
Total	3,862

We aimed to understand how the visions and goals of the HAI approaches relate to the increasing inequalities in capitalism (Piketty, 2020) in the published content of these university departments. We believe that such educational institutions should strive to analyze the topics below, elaborate on possible solutions, and redress the grievances that harm people globally.

Capitalism: the universal system that prioritizes “the endless accumulation of capital” (Wallerstein, 2004, p. 24), sustains or deepens crises for specific societies or communities, especially in the Global South (Böröcz, 2009; Piketty, 2014, 2020). We aim to understand how the human-centered approach relates to capital accumulation and vital wealth inequalities emerging at sub- and supranational levels.

Vulnerable groups: it refers to any group harmed by racism, sexism, gender bias, xenophobia, nativism, antisemitism, islamophobia, economic, or environmental crises. Marginalized groups such as the Global South, LGBTQ communities, religious and ethnic minorities, refugees, immigrants, low-income citizens, people with physical or cognitive disabilities also belong to these circles (Fountain, 2021). The theme also includes the direct juxtaposition of the elite's interest versus the exposed ones.

Power issues (decision-making): making decisions over people's lives without asking their opinion about the implemented policies by governments or tech corporations. The theme entails decision-makers' activities, including controlling AI by measures or abusing power by surveillance techniques based on AI-driven technology, such as the Chinese "scoring system," (Starke & Lünich, 2020).

Supportive attitudes: according to Hagendorff's (2020) argument, these subthemes do not emerge remarkably in the "Technical solutions" category within AI ethics:

a) Social responsibility: policies that address societal challenges, public engagement activities, charitable giving, and efforts to benefit a sustainable environment.

b) Welfare: this subtheme is not axiomatic as it deals with a severe dichotomy; thus, we analyzed whether (1) the connection of AI and welfare refers to sustaining (welfare's) status quo and disregarding policies on decreasing inequalities or (2) it analyzes how AI could ease major wage and redistribution cleavages between classes and regions.

c) Help: this subtheme includes any effort aiding marginalized communities, people in need, or fighting against climate change.

d) Care & nurture: any activity related to healthcare or arguments on necessary equipment and innovations in medical treatment and taking care of young children to keep their systems developing and healthy.

Technical solutions: Hagendorff's argument is based on Gilligan's claims (1982), that is, male-dominated justice ethics is calculating, rational, and logic-oriented, mostly disregarding the ethics of empathic and emotion-oriented care:

e) Safety: avoiding AI "side-effects," such as harmful multi-agent approaches, uncertainty, hacking, accidents in machine learning systems are parts of this subtheme (Amodei et al., 2016).

f) Robustness: building reliable and secured machine-learning systems is a crucial area of AI studies.

g) Explainability: this subcategory implies arguments on how and why an artificial intelligence algorithm makes decisions while it preserves its accuracy.

h) Accountability: it is a closely related concept to transparency, which needs transparent processing operations. In short, accountability can be considered as vital data protection and privacy emphasizing principle (Vedder & Naudts, 2017).

i) Privacy: machine-learning capabilities can put privacy and data protection at risk. Consequently, any text unit that argues the issue of privacy is relevant to this theme.

j) Responsibility: one of the essential questions in AI techniques and machine learning capabilities considers the entity which is responsible for (1) programming the algorithm, (2) controlling its functions, (3) and taking responsibility for harmful happenings, for instance, accidents, caused by AI-driven programs.

k) Non-maleficence: causing no harm should be vital for planning and operating AI-driven technologies. This theme falls under analyzing possible harmful physical and psychological effects and preventing measures.

l) Justice & fairness: on the one hand, AI-driven technology should be controlled by law and order to rightfully involve the court of justice if any non-compliant activity is perceived. On the other hand, unfairness might refer to the fact that AI-based calculations obstruct several marginalized groups from receiving loans and medical care, distracting the opportunity to provide fair services by avoiding gender, race, and financial bias.

m) Transparency: it is an essential element of efficient accountability frameworks by ensuring that an algorithmic process is observable and information considering future behavior is supplied (Alhadeff et al., 2012).

RESULTS AND ANALYSIS

How the institutes define HAI

First, we provide the descriptions of the institutes' definitions on human-centered artificial intelligence to overview their approaches on this research area and seek possible connections between the visions.

The Utrecht University mostly stresses personalization; therefore, it describes HAI as a developing technique that understands and predicts human choice behavior and convinces people to make efficient and environment-friendly decisions, including intelligent interactive information systems and personalized interaction to maximize user satisfaction. This interpretation regards HAI as a product that is designed to be sold and foster convenience.

The University of Maryland imagines a possible, alternative future filled with devices that dramatically amplify human abilities, empower people, and ensure human control. This institution considers HAI as a tool designed for the people but avoids mentioning business interests and gaining profit. According to the institute's vision, HAI enables people to perceive, create, think, and act by combining user experiences with embedded AI support services that users desire.

Stanford University claims that HAI develops frameworks representing different stakeholders focusing on interdisciplinary collaboration in AI design, development, and management. Stanford University's explicit vision on HAI is to develop a tool that fosters a better future for humanity.[2] Therefore, the research institute argues that AI's designer team must comprise humanity's broad representatives. It claims that the creators of AI have a collective responsibility to guide machine-learning approaches in an ethical way, that is, fostering positive effects on the globe. This research institute declares that it aims to help future leaders prepare to "learn, build, invent and scale with purpose, intention and a human-centered approach."[3] On

[2] The Stanford University's Human-Centered Artificial Intelligence research institute emphasizes its vision on HAI briefly via its title page under the section of "Advancing AI research, education, policy, and practice to improve the human condition"(see at https://hai.stanford.edu/) and discusses it in detail under the "About" section (see "Welcome to the Stanford Institute for Human-Centered Artificial Intelligence – Letter from the Denning Co-Directors" at https://hai.stanford.edu/about/welcome). Date accessed: 12 January 2022.

[3] See "Welcome to the Stanford Institute for Human-Centered Artificial Intelligence – Letter from the Denning Co-Directors" at https://hai.stanford.edu/about/welcome. Date accessed: 12 January 2022.

the contrary, the broad access of ordinary people to AI techniques is missing from the statement above because it focuses on an AI designing process aligned with “optimistic techno-scientific visions” (Dandurand et al., 2020, p. 600).

Northwestern University defines HAI as a socio-technical system to advance decision-making and “creative and analytical thinking, feeling, and doing.” The institute’s introduction highlights that machine-learning and data-mining approaches are to augment “human emotion, cognition, and behavior.”

The University of Bologna emphasizes that machine-learning approaches are helpful in the fight against organized crime, cyberbullying, cyber-crimes, fake news, and hate speech. The Italian university claims that HAI techniques are necessary to resist criminal activities. In other words, the University of Bologna defines HAI as a sufficient tool to fight against these phenomena. In contrast, the University of Bologna does not define how AI-driven techniques could prevent the proliferation of the challenges above.

In a nutshell, every institution has a unique approach to HAI but lacks universal values except one: they define HAI in relation to the people. These departments primarily suggest that HAI, in some ways, is adjusted to people’s needs. This is an important finding because the departments above suggest, in a very diverging way, though, that HAI is mainly for all of humankind and not for profit accumulation (Wallerstein, 2004). Although, it is important to mention that Utrecht University, to a certain extent, tends to consider AI as a product. Critically, however, these definitions have vital limitations. For example, the definitions treat humanity as a homogenous mass, sharing universal goals, needs, and interests. Even though people are born with equal dignity and have the right to happiness, we argue that humanity is not homogenous: diverging people and regions struggle mostly with different and sometimes overlapping challenges. Therefore, we argue that AI and ethics attached to the design process can be humanistic if adjusted sensitively to “individual” situations (Hagendorff, 2020). In addition, as we will show later, the institutes often highlight challenges for the local marginalized communities but mostly disregard problems proliferating in the Global South, such as the lack of water supply, starvation, diseases, and wars that affect a significant part of Africa and a large part of Asia and South America.

General outcomes

We introduce our in-depth analysis with specific examples to give insights into our main arguments starting with the relation of capitalism and the visions on AI. Even though our analysis is qualitative, we aim to present the detailed findings of the theme proportions, which helps us compare the institutions’ agenda setting on the HAI imaginary and its bonds to the topics above (see Table 2-4).

Table 3 – Crosstab with themes, column percentages based on the number of documents in which themes appear (row “N= documents”)

	University of Bologna	**Utrecht University**	**University of Maryland**	**Northwestern University**	**Stanford University**
1) Capitalism	0	0	0	0	2,8%
2) Vulnerable groups	3,1%	16,9%	9,6%	25,0%	34,1%
3) Power issues (decision-making)	6,3%	27,1%	7,4%	13,9%	48%

4) Supportive attitudes	34,4%	42,4%	23,4%	27,8%	72,2%
5) Technical solutions	12,5%	39%	14,9%	16,7%	58%

We coded numerous paragraphs (n=3,862) and perceived that the most frequent theme is Supportive attitudes (n=1,704) followed by Technical solutions (n=1,049), Power issues (n=673), Vulnerable groups (n=420), and Capitalism (n=16).[4] Table 3 presents the aggregated results of themes and subthemes and suggests that the Supportive attitudes category is the most frequently emerging theme in every institute's published content. Stanford University publishes the most content as it provides almost four times more texts than the second most "productive" college, namely the University of Maryland. Additionally, Stanford University produces more content than the other institutions together. Stanford University covers every theme and subtheme, while the University of Bologna is the least diverse in terms of themes. The most productive college dominates every theme and most of the subthemes with three exceptions: Social responsibility & sustainability emerges with larger shares on the two European universities' websites, and Explainability and Transparency occur with a higher share in the Utrecht University's online content than on Stanford University's webpage (see Table 4).

Table 4 – Crosstab with themes and subthemes, column percentages based on the number of documents in which themes and subthemes appear

	University of Bologna	Utrecht University	University of Maryland	Northwestern University	Stanford University
1) Capitalism	0	0	0	0	2,8%
2) Vulnerable groups	3,1%	16,9%	9,6%	25,0%	34,1%
3) Power issues (decision-making)	6,3%	27,1%	7,4%	13,9%	48,0%
4) Supportive attitudes	-	-	-	-	-
a) Social responsibility & sustainability	28,1%	30,5%	2,1%	19,4%	23,3%
b) Welfare	0	1,7%	1,1%	2,8%	25,6%
c) Help	0	1,7%	14,9%	2,8%	48,3%
d) Care & nurture	9,4%	22,0%	10,6%	11,1%	49,4%
5) Technical solutions	-	-	-	-	-
e) Safety	6,3%	23,7%	10,6%	8,3%	27,3%
f) Robustness	0	1,7%	2,1%	0	8,5%
g) Explainability	3,1%	8,5%	1,1%	2,8%	3,4%
h) Accountability	3,1%	1,7%	0	0	8,5%
i) Privacy	3,1%	6,8%	10,6%	0	18,5%
j) Responsibility	0	5,1%	1,1%	0	11,6%
k) Non-maleficence (causing no harm)	0	3,4%	0	5,6%	7,7%
l) Justice & fairness	6,3%	18,6%	2,1%	8,3%	33,2%
m) Transparency	3,1%	11,9%	0	0	9,4%

[4] To overview these results, see Table 2.

An important outcome is that Supportive attitudes emerge more frequently than Technical solutions. In contrast to Hagendorff's (2020) results, which shows that the Technical solution theme dominates ethics guidelines, the five analyzed universities rather focus on the Supportive attitudes category than overemphasize the "male-dominated" approach. This outcome suggests that different agenda-setting can be perceived if one contrasts AI ethical guidelines to university departments and labs' agenda on HAI.

Specific analysis

We introduce our findings in detail on the five themes, starting with Capitalism and finishing with Technical solutions. Moreover, we characterize three subthemes from Supportive attitudes: (1) Social responsibility & sustainability, (2) Welfare, and (3) Care & nurture because these topics are discussed much more in detail than Help, which is a rather general expression without concrete, programmatic guidelines and suggestions. We scrutinize the three subthemes above because we argue that the Covid-19 pandemic, the ecological crisis, and flourishing inequalities are intertwined and should be analyzed, if not eased, as soon as possible by the opportunity that artificial intelligence offers us to consider it as "human-centered." Note that we do not analyze Technical solutions' subthemes but the theme as a whole because we aim to compare the five main themes to supply an easy-to-follow analysis and focus on the bigger picture to avoid being lost in detail.

Capitalism

A striking result is that only Stanford University considers the theme of capitalism worth mentioning when introducing the visions, goals, and expectations on artificial intelligence. The rest of the institutions lack any argument on global capitalism, its ties to artificial intelligence, and the institutions' relation to the prevailing economic and political system. In turn, Stanford introduces the pros and cons of capitalism by publishing interviews with persons who support or criticize capital accumulation. A postdoctoral research fellow at Stanford University analyzes the ties between capital accumulation and his institute:

> Stanford certainly has the institutional capital and cultural cachet to influence the AI industry; the question is how it will use that power. The major problems of the 21st century are problems of distribution, not production. There's already enough to go around; the problem is that a small fraction of humanity monopolizes the resources. In this context, making AI more 'human-centered' requires focusing on the problems facing the majority of humanity, rather than Silicon Valley.
>
> To pioneer a human-centered AI R&D agenda, thought leaders at Stanford's HAI and elsewhere will have to resist the powerful incentives of global Capitalism and promote things like funding AI research that addresses poor people's problems; encouraging public participation in decision making about what AI is needed and where; advancing AI for the public good, even when it cuts into private profits; educating the public honestly about AI risks; and devising policy that slows the pace of innovation to allow social institutions to better cope with technological change. (Miller, 2020a)

The argument above has several important implications. First, it stresses that Stanford, with an endowment of more than $15 billion, which places the university among the top four colleges in the United States (Piketty, 2014, p. 447), has enough resources to develop AI-driven techniques that would improve people's well-being globally. Second, it brings attention to a choice that must be made sooner or later from the institute's side: does Stanford develop strategies that might ease global problems or join corporations that chase profit? Finally, the answer above juxtaposes the economic elite to the "common" people and suggests that academics have the role of encouraging the masses to participate in decision-making and informing citizens about AI techniques in detail. Even though the claim above is critical about capitalism, only a few criticisms emerge on Stanford's web page, and most of the content do not touch upon criticism of social and business arrangements of capital accumulation.

Vulnerable groups

Vulnerable groups appear in every institution's content, but only certain marginalized circles are typically mentioned. Most of the analyzed texts focusing on HAI imaginary related to vulnerable groups consider marginalized circles mostly locally but not globally. This is problematic because the Global South and its vast, struggling masses are underrepresented in the content arguing the visions and goals of HAI.

Considering the ethical questions emerging within the topic of machine-learning algorithms, the University of Utrecht explicitly addresses the problem of biased programming of AI-driven techniques:

> Although computers are often advertised as objective and neutral, the way in which the computers are 'raised' provokes questions. Doubts arise on whether or not the current anti-discrimination laws are well-equipped enough to deal with this and if they provide the necessary safeguards (University of Utrecht, 2019).

The Dutch university suggests that artificial intelligence is far from neutral because it is created by humans who may have stereotypes, be pressured in design processes, lack empathy, or do not care about the potential adverse impact of AI on marginalized groups. Sadly, the analyzed contents do not explain how vulnerable groups should be defended from biases.

Stanford University has the highest percentage of texts of all the institutions dealing with vulnerable groups. In turn, it acknowledges that the research field of AI and academia generally have not yet reflected diversity issues to the necessary extent, and it will take time to change such systemic problems. Even though Stanford University deals the most with vulnerable groups, it is important to emphasize that it mainly mentions vulnerable groups living within the United States but rarely highlights other marginalized groups' problems and the possible solutions, such as ceasing starvation, poverty, and life-threatening jobs for the majority residing in the Global South. The example below shows an essential but local problem, which is Hispanic people's, black communities', and women's underrepresentation in American history textbooks used in Texas:

> The most dramatic finding in the Texas history project was the virtual absence of Hispanic people, who received almost no attention outside of the Mexican-American War. Women fared better, but they too were discussed far less frequently than men. (Andrews, 2020)

The HAI institutes' web pages mainly concentrate on the Global North and its challenges in an era when AI "should be built so as to have net benefits for the whole of society" (Baum, 2017, p. 544) that could contribute to bettering the lives of poorer societies, such as beneficence, non-maleficence, autonomy, justice, explicability, safety, and early disease detection (Berberich et al., 2020; Floridi & Cowls, 2021). Among others, we argue that water supply is a key segment of redressing vulnerable groups' grievances, especially in the Global South. The sufficient supply of drinking water by AI is a vital opportunity to support the survival of the most vulnerable ones and prevent several fatal or non-fatal diseases. Artificial intelligence's role in supporting sufficient water supply implies repairing eroded equipment, analyzing water quality, and detecting inhabited areas without drinking water. Stanford University mentions problems with water supply in the Global South, however, only to a minimal degree. For example, there are instances where it is discussed that satellites and AI techniques might augment each other and foster mapping African countries' poor infrastructure. Consequently, constructing water-supplying pipelines could be developed much more precisely due to the rich data analyzed by AI-driven software.

Power issues

The theme of Power issues has close ties to legislation and measures on artificial intelligence decision-making. Every institute published content on power issues but on a very different scale: the University of Bologna focused to the smallest, while Stanford University to the largest extent on this theme.

Interestingly, large tech companies' power, that is, collecting, sharing, or exploiting user data for commercial or political goals, emerges to a different extent in the analyzed institutes' content. Even though the University of Bologna's education program implements the intersection of AI and business, we did not find any evidence of criticizing big tech's power supported by AI on the institute's website.

Texts published via Stanford University's HAI website have different approaches to power issues and large tech corporations with mainly a common aspect: accountability. Some of these articles argued that laws initiated by the U.S. government must regulate big tech companies; others suggested that companies should regulate themselves. Cathy O'Neill, the author of Weapons of Math Destructions, argued in an interview conducted at Stanford University that three diverging aspects could be perceived as related to power imbalance and accountability, namely in the (1) United Kingdom and "Europe," (2) China, and (3) the United States. In particular, the following argument reflects on facial recognition systems' unethical coded bias:

> … in the States, we live in the wild, wild west. We are home to these tech companies and yet don't have meaningful regulations. Arguably there are more laws that govern my behavior as an independent filmmaker trying to get broadcast on PBS than govern Facebook where a billion people go for their information and political speech (Miller, 2020b)

The interviewee suggests that the United States' federal government should regulate facial recognition software by laws, and these measures must rely on transparent guidelines to balance big tech's power, which creates these surveillance programs to avoid prejudice.

The Utrecht University joins the argument by emphasizing the role of ethical designing in responsible, autonomous systems:

> Increasingly, computer systems with some degree of autonomy are being employed in practice. Such artificially intelligent software can do things that, when done by humans, are regulated by law. For example, self-driving cars have to obey the traffic laws, online information systems have to comply with data protection law, care robots can damage

> property or the health of the persons they care for, and autonomous weapons have to comply with the laws of war (University of Utrecht, 2021)

Even though GDPR-regulations started within the EU and the United Kingdom, the Utrecht University goes further and elaborates on guidelines for governments – they did not outline which governments they refer to – that should use mobile phone data to design effective measures during the proliferating pandemic. Privacy concerns, however, arise, and the institution suggests that data anonymization could protect citizens' privacy. Besides the warning above, there is no criticism reflecting upon big tech corporations or any harmful consequences of their AI-driven techniques.

The University of Maryland mostly seeks sponsorships and collaborations with tech companies via its website, but we also perceived content criticizing firms' biases towards women. One of the submitted abstracts of a speaker series argues that big tech companies "create a work environment of bias, hostility and devalue"; therefore, fewer women worked for tech companies in 2014 (25% of the employees) than in 1990 (31%) and female's quit rate is also higher than men's deliberative decision to leave these firms.

Finally, Northwestern University claims that the HAI institute provides "rigorous research insights to industry and government leaders – contributing to future products from international technology companies", but lacks the critical approach to scrutinizing big tech corporations' ties to power issues.

Supportive attitudes: Social responsibility & sustainability, Welfare, and Healthcare

We continue our analysis with the theme of Supportive attitudes in which Social responsibility & sustainability is addressed with similar shares except for the University of Maryland, which hardly mentions the issue. Even though Social responsibility & sustainability regularly emerges within the University of Bologna's website, it is not argued how and when AI would contribute to societal questions or ecology-saver policies. Besides, the Italian institute outlines what should research on AI focus on and disregards the implementation of the technique. Utrecht University argues that societal issues and sustainability should be analyzed together because irrigating crops optimally, storing renewing energy, and fighting local air pollution are interconnected issues. The vital feature of the argument above is that Utrecht University aims to transform knowledge on the ecosystem by being "open to the outside world." Unfortunately, it did not express when sufficient knowledge would be available and how the recommendations based on these findings should be implemented. Stanford University suggests that the knowledge captured by AI should be translated into "community-based" decision-making processes as soon as possible. According to Stanford University's articles, AI is useful because it can process ecological problems in multiple dimensions, whereas the human brain cannot absorb so much information. In turn, Northwestern University highlights another aspect of societal challenges and social responsibility:

> We face a global demand for new ways of continuously training and reskilling workers, and we need new socio-technical systems to better enable and advance human sensemaking, decision-making, creative and analytical thinking, feeling, and doing. New techniques are needed that integrate artificial intelligence, machine learning and data-mining approaches in the service of augmenting human emotion, cognition, and behavior. (Northwestern University, 2021)

Northwestern University addresses the problematic issue of the relation between human and non-human labor force. It suggests that robots and software should be adjusted to fill gaps or difficulties for human laborers. Although Northwestern does not argue recommendations in detail, it collects research papers on the aforementioned challenge (Hong et al., 2020; Zacks & Franconeri, 2020). Professor Susan Athey, an associate director of HAI institute at Stanford University, also brought attention to the role of AI in the labor market: machine-learning approaches should augment more than replace human workers. Additionally, she argues that there are many tools to evaluate data to help displaced workers overcome difficulties if they lose their jobs due to automation. She suggests that finding upskilling courses that suit displaced workers might foster the solution.

The question of welfare barely emerges in four departments' contents. In contrast, Stanford University emphasizes this theme more than others (see Table 4). Probably the most exciting argument on welfare and AI is based on the following perception articulated by Sucheta Ghoshal, an assistant professor at the University of Washington who introduces India's case: "It [AI] was presented as supporting a welfare pipeline but ended up being a massive surveillance and security risk used for religious/caste segregation. (Waikar, 2021)

The statement above is important because Ghoshal highlights vital issues. First, her claim refers to Aadhaar, India's large-scale biometric identification system. The argument brings attention to a Global South country that could have benefitted from artificial intelligence but misses the chance to reduce inequality by machine-learning technologies. The biometric identification system, which covers more than 1.2 billion people, was advocated by the Indian government claiming that Aadhaar would reduce fraud and allow the poor to reach more subsidies. Aadhaar, which suffers from several glitches such as network outages, is linked to food subsidies, pensions, medical reimbursements, and disaster emergencies. If there is a fault within Aadhaar, which, according to The Guardian's report, regularly emerges, access to subsidies may suddenly stop (Ratcliffe, 2019). Even though surveillance is unethical and can be one of the cornerstones of oppression, we have a different reason to bring attention to the case above. Besides its observatory nature, an error within Aadhaar can be fatal if food subsidies are limited or banned by a bug emerging in the system. In several instances, Aadhaar did not function correctly, and people died due to starvation (Ratcliffe, 2019). We argue that besides surveillance, the fatal consequences of malfunctioning should have been presented in the analyzed documents since famine also poses a threat to the poor, probably in a more serious way than observation. In other words, famine is a more severe problem than observation, but Ghoshal, aligned with "Western" values, emphasizes the latter but does not focus on the former.

Finally, we introduce another essential problem within the intersection of Supportive attitudes and AI ethics: prejudice in healthcare. The most attractive example was provided by Stanford University that considers the role of AI in healthcare as a predicting algorithm, which might suffer from serious prejudice if vulnerable groups' unmet needs for treatment are not resolved:

> For example, a hospital might use its electronic healthcare records to predict which patients are at risk of cardiovascular disease, diabetes or depression and then offer high-risk patients' special attention. But women, Black people, and other ethnic or racial minority groups might have a history of being misdiagnosed or untreated for these problems. That means a predictive model trained on historic data could reproduce historical mistreatment or have a much higher error rate for these subgroups than it does for white male patients (Miller, 2020c)

Stanford University rightfully argues that marginalized groups could keep being mistreated based on AI suggestions – even if the program was created by the best intent – if redistribution remains unfair and the former data is biased.

Technical solutions

Our results suggest that three subthemes are salient within technical solutions: safety, privacy, and justice/fairness. The issues of safety and privacy are regularly intertwined in the analyzed texts, which argue that decision-makers aim to install policies on artificial intelligence and machine-learning algorithms to control big tech companies' endeavors of abusing personal data. However, as Jessica Fjeld argues at Stanford's AI & Human Rights Symposium, the problematic part with governmental regulations is that "offloading of liability onto machines may benefit only the corporations that make those machines, and not society in general." For instance, AI-driven surveillance techniques could be biased, harming mostly Latinos and black communities. Several governments are willing to buy or design these programs and even outline what type of data they aim to collect by surveillance systems. This is the point where justice and fairness kick in. What happens if the program makes false labels and predictions? It will deepen the crisis of societally vulnerable people. The aforementioned problem is aligned with Monaham's argument (2021), by which he claimed that decolonizing surveillance and relevant studies could challenge Western white male supremacy.

Bias in design has emerged as a relevant topic within the AI sphere (Metz, 2021).This is often linked to the fact that technical solutions are still designed mainly by white men. If we look at the web pages of the analyzed institutions, although certain diversity is present – especially in the case of Standford University – there is still a dominance of white men. Regulation that supports more diverse research teams across the Global South is inevitable if the HAI concept is geared towards creating technical solutions that address grievances rather than maintaining a status quo steeped in racism. Most of the analyzed institutions admit explicitly or implicitly that the current exclusionary nature of technical solutions is unaffordable. In turn, no specific steps have been presented that might challenge the biased-led AI industry.

CONCLUSION

In this book chapter, we analyzed five research institutes' official web pages having "Human-centered" and "AI" expressions in their names to scrutinize their published content on the HAI approach's visions on humanity and ethical values connected to capitalism, vulnerable groups, power issues, supportive attitudes, and technical solutions. Every analyzed research institute is from the Global North, and among many questions, we investigated how they outline the relation of AI-driven techniques to local and global communities. These institutes define the HAI approach differently except for one thing: they outline their definitions related to humankind, which they consider as a homogenous mass with the same needs. We argue that different groups have diverging needs, but, of course, specific needs, like well-being, can be overlapping.

The five analyzed research institutes bring attention to supportive attitudes rather than technical solutions in their published content. This is an important result because former research highlighted a reversed outcome when ethical guidelines from the non-academic sectors were analyzed (Hagendorff, 2020). In other words, the scrutinized institutes recognize that they should put supportive attitudes in the center of their published content rather than technical documents to demonstrate their humanistic efforts. Stanford University is a salient institute as it published more content than the others and covered every topic we analyzed. Even though Stanford University made the most effort to introduce the visions of HAI ethics, it fails to provide concrete plans on redressing grievances on a global level, resolving surging inequalities in capitalism, and fighting against worldwide racism.

Researchers argue that there is a need for constructing bridges between AI ethics and its implementations into technical solutions (Hagendorff, 2020). All institutes acknowledge that they collaborate with tech companies to a certain extent. However, they do not explain how these co-operations will help reduce inequalities, ease starvation, provide water supply, detect the lacking infrastructure, and refine biased recognition systems. Detailed, understandable, and transparent explanations are crucial to acquire the trust from the public, especially from laborers whose jobs are on the line due to the rapid automatization. We argue that transparent collaborations could foster plans for reducing poverty in the Global North and the Global South because corporations or other organizations could join this effort with their resources and know-how to implement their ethical principles into practice. Until these co-operations are not transparent, none of the stakeholders will be motivated to go beyond ethics discourses that operate only as an assurance for the public and investors (Kerr et al., 2020).

The institutes above claim that they collaborate with tech firms and tend not to criticize capitalism or Western neoliberal values (Stark et al., 2021). We found a few paragraphs at Stanford University's HAI website where fair and universal redistribution appears as a pivotal need within capitalist production. We argue that much more content should be published on reducing inequalities. We suggest that the other institutes, along with Stanford University, should make greater efforts together to analyze these challenges and create plans for easing local and global societal challenges. As these institutes are parts of the wealthy countries in the center (Wallerstein, 2004), they have the most extensive resources to help people in need.

One of the ethical values that most institutes emphasize is human responsibility in designing. Some institutes claim that excuses based on the algorithms' neutrality are not defendable because bad designs, mostly affected by poor or the absence of ethics, derive from human errors (Greene et al., 2019). The perspective above is important because it suggests that beyond the aim of ethics-washing in self-regulation (Bietti, 2020), philosophers, social scientists, and citizens should be diversely involved in the designing processes. We agree with researchers who argue that building machine-learning systems should be built after profound consultations with citizens with the intention of "understanding of users' characteristics, the methods of coordination, the purposes and effects of an intervention; and with respect for users' right to ignore or modify interventions" (Cowls et al., 2019, p. 19).

As we detected in our analysis, some institutes (primarily Stanford University) introduced that very narrow groups (e.g., engineers) create AI-driven software. Additionally, we found many articles in which the HAI approach appeared concerning local minorities and biases haunting their everyday lives. Why do we emphasize these observations? Extant research proves that tech companies are not motivated to extend their attention beyond these circles and ease the severe challenges of struggling people (Washington & Kuo, 2020). These corporations' main aim is profit accumulation in the center; therefore, they act globally (sell their products anywhere) and think locally (keep profit in the centrum). In turn, HAI research institutes think globally by highlighting their awareness of humanitarian crises beyond the Global North but mostly focusing on local marginalized communities' disadvantages. The Global South and its numerous problems are underrepresented in the ethical and humanistic visions of AI. How can we call it "Human-centered AI" if these problems remain unsolved both in communication and practice? Big tech corporations and institutes from the Global North have the necessary resources and knowledge to utilize ethical practices to redress diverging grievances together in different regions. One of the biggest challenges in such a helpful collaboration is stressed by Stark and colleagues: “how can members of diverse communities, often with asymmetric access to wealth and power, work together to ensure justice, equality, and fairness exist not just in principle but also in practice” (Stark et al., 2021, p. 273). Our findings suggest that the essential question above remained unanswered in detail. Nonetheless, we think that solutions adjusted to the most severe grievances cannot be redressed without implementing struggling communities’ will to help in democratic and ethical ways.

FUNDING

The study was funded by the Hungarian NRDI Office, grant number OTKA K 142207.

REFERENCES

Alhadeff, J., Van Alsenoy, B., & Dumortier, J. (2012). The accountability principle in data protection regulation: origin, development and future directions. In D. Guagnin, L. Hempel, C. Ilten, I. Kroener, D. Neyland, & H. Postigo (Eds.), Managing privacy through accountability (pp. 49-82). Palgrave Macmillan. https://doi.org/10.1057/9781137032225_4

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

Andrews, E. L. (2020). Whose History? AI Uncovers Who Gets Attention in High School Textbooks. Retrieved 5 October 2021, from https://hai.stanford.edu/news/whose-history-ai-uncovers-who-gets-attention-high-school-textbooks

Baum, S. D. (2017). On the promotion of safe and socially beneficial artificial intelligence. AI & SOCIETY, 32(4), 543-551. https://doi.org/10.1007/s00146-016-0677-0

Berberich, N., Nishida, T., & Suzuki, S. (2020). Harmonizing Artificial Intelligence for Social Good. Philosophy & Technology, 33(4), 613-638. https://doi.org/10.1007/s13347-020-00421-8

Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain. https://doi.org/10.1145/3351095.3372860

Böröcz, J. (2009). The European Union and Global Social Change: A Critical Geopolitical-Economic Analysis (1 ed.). Routledge. https://doi.org/10.4324/9780203873557

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Cowls, J., King, T., Taddeo, M., & Floridi, L. (2019). Designing AI for social good: Seven essential factors. Available at SSRN 3388669.

Dandurand, G., Claveau, F., Dubé, J.-F., & Millerand, F. (2020). Social Dynamics of Expectations and Expertise: AI in Digital Humanitarian Innovation. Engaging Science, Technology, and Society, 6, 591-614.

Demeter, M. (2020). Academic Knowledge Production and the Global South. Questioning Inequality and Under-representation (1 ed.). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-52701-3

Floridi, L., & Cowls, J. (2021). A Unified Framework of Five Principles for AI in Society. In L. Floridi (Ed.), Ethics, Governance, and Policies in Artificial Intelligence (pp. 5-17). Springer International Publishing. https://doi.org/10.1007/978-3-030-81907-1_2

Fountain, J. E. (2021). The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. Government Information Quarterly. https://doi.org/10.1016/j.giq.2021.101645

Gilligan, C. (1982). In a diferent voice: Psychological theory and women's development. . Harvard University Press.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. Proceedings of the 52nd Hawaii international conference on system sciences,

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. Proc. ACM Hum.-Comput. Interact., 4(CSCW1), Article 068. https://doi.org/10.1145/3392878

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. Big Data & Society, 7(1), 2053951720915939. https://doi.org/10.1177/2053951720915939

Metz, V. (2021). Who Is Making Sure the A.I. Machines Aren't Racist? Retrieved 19 December 2021, from https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html

Miller, K. (2020a). HAI Fellow Colin Garvey: A Zen Buddhist Monk's Approach to Democratizing AI, Retrieved 5 October 2021, from https://hai.stanford.edu/news/hai-fellow-colin-garvey-zen-buddhist-monks-approach-democratizing-ai

Miller, K. (2020b). Coded Bias: Director Shalini Kantayya on Solving Facial Recognition's Serious Flaws. Retrieved 5 October 2021, from https://hai.stanford.edu/news/coded-bias-director-shalini-kantayya-solving-facial-recognitions-serious-flaws

Miller, K. (2020c). When Algorithmic Fairness Fixes Fail: The Case for Keeping Humans in the Loop, Retrieved 5 October, 2021, from https://hai.stanford.edu/news/when-algorithmic-fairness-fixes-fail-case-keeping-humans-loop.

Monahan, T. (2021). Reckoning with COVID, racial violence, and the perilous pursuit of transparency. Surveillance & Society, 19(1), 1-10.

Northwestern University (2021) Human-centered AI. Retrieved 5 October 2021, from https://hci.northwestern.edu/research/human-centered-ai.html

Patton, M. Q. (1990). Qualitative evaluation and research methods, 2nd ed. Sage Publications, Inc.

Piketty, T. (2014). Capital in the Twenty-First Century. Belknap Press.

Piketty, T. (2020). Capital and Ideology. Harvard University Press.

Ratcliffe, R. (2019). How a glitch in India's biometric welfare system can be lethal. Retrieved 19 December 2020, from https://www.theguardian.com/technology/2019/oct/16/glitch-india-biometric-welfare-system-starvation

Rooduijn, M. (2014). The mesmerising message: The diffusion of populism in public debates in Western European media. Political Studies, 62(4), 726-744.

Stark, L., Greene, D., & Hoffmann, A. L. (2021). Critical Perspectives on Governance Mechanisms for AI/ML Systems. In J. Roberge & M. Castelle (Eds.), The Cultural Life of Machine Learning (pp. 257-280). Palgrave Macmillan. https://doi.org/https://doi.org/10.1007/978-3-030-56286-1_9

Starke, C., & Lünich, M. (2020). Artificial intelligence for political decision-making in the European Union: Effects on citizens' perceptions of input, throughput, and output legitimacy. Data & Policy, 2, e16, Article e16. https://doi.org/10.1017/dap.2020.19

University of Utrecht (2019). The lunch meeting on algorithms and diversity, Retrieved 5 October 2021, from https://www.uu.nl/en/news/the-lunch-meeting-on-algorithms-and-diversity.

University of Utrecht (2021). AI, Ethics and Law. Retrieved 5 October 2021, from https://www.uu.nl/en/research/human-centered-artificial-intelligence/special-interest-groups/ai-ethics-and-law

Tiyasha, Tung, T. M., & Yaseen, Z. M. (2020). A survey on river water quality modelling using artificial intelligence models: 2000–2020. Journal of Hydrology, 585, 124670. https://doi.org/https://doi.org/10.1016/j.jhydrol.2020.124670

Vedder, A., & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. International Review of Law, Computers & Technology, 31(2), 206-224. https://doi.org/10.1080/13600869.2017.1298547

Wallerstein, I. (2004). World-system Analysis – An Introduction. Duke University Press.

Waikar, S. (2021). Designing Anti-Racist Technologies for a Just Future, Retrieved 5 October 2021, from https://hai.stanford.edu/news/designing-anti-racist-technologies-just-future

Washington, A. L., & Kuo, R. (2020). Whose side are ethics codes on? power, responsibility and the social good Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain. https://doi.org/10.1145/3351095.3372844

Xiang, X., Li, Q., Khan, S., & Khalaf, O. I. (2021). Urban water resource management for sustainable environment planning using artificial intelligence techniques. Environmental Impact Assessment Review, 86, 106515. https://doi.org/https://doi.org/10.1016/j.eiar.2020.106515

Zacks, J. M., & Franconeri, S. L. (2020). Designing Graphs for Decision-Makers. Policy Insights from the Behavioral and Brain Sciences, 7(1), 52-63. https://doi.org/10.1177/2372732219893712

BIONOTES

Lilla Vicsek, Ph.D., is Associate Professor at Corvinus University of Budapest. Her work for over a decade has focused on issues related to science and society, with publications appearing in journals such as Science as Culture, New Genetics and Society, Science Communication, and Journal of Sociology. Currently, her main research focus is the constitutive role of expectations regarding artificial intelligence. She is especially interested in how these expectations are related to ethics and power issues.

Tamás Tóth is an Assistant Professor at the University of Public service. His research interest includes populist political communication styles, journalism studies, and academic knowledge production. Recently, he elaborated the content analysis refinements of explicit and implicit populism to scrutinize manifest and latent dichotomies in populist political communication. He published in journals such as the International Journal of Communication, the Journal of Contemporary European Studies, the European Journal of Science and Theology, and Scientometrics.