Sustainable AI

An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence

Larsson, Stefan; Anneroth, Mikael; Felländer, Anna; Felländer-Tsai, Li; Heintz, Fredrik; Cedering Ångström, Rebecka

2019

Document Version:
Publisher's PDF, also known as Version of record

An inventory of the state of knowledge of ethical, social,
and legal challenges related to artificial intelligence

SUSTAINABLE AI

Stefan Larsson, Lund University, Fores
Mikael Anneroth, Ericsson Research
Anna Felländer, AI Sustainability Center
Li Felländer-Tsai, Karolinska institutet
Fredrik Heintz, Linköping University
Rebecka Cedering Ångström, Ericsson Research

Bibliometric analysis by Fredrik Åström, Lund University.

This report is based on the Vinnova-funded project
Hållbar AI – AI Ethics and Sustainability, led by Anna Felländer.

AI SUSTAINABILITY CENTER

SUSTAINABLE AI

An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence

Stefan Larsson, Lund University, Fores
Mikael Anneroth, Ericsson Research
Anna Felländer, AI Sustainability Center
Li Felländer-Tsai, Karolinska institutet
Fredrik Heintz, Linköping University
Rebecka Cedering Ångström, Ericsson Research

AI Sustainability Center, 2019

THE AI SUSTAINABILITY CENTER is a multidisciplinary center for responsible and purpose-driven technology, based on Nordic values. It brings together actors from the business sector, the public sector and other non-governmental organisations, as well as experts from various academic fields, in a collaborative initiative for piloting and implementing AI sustainability strategies and frameworks.

AI SUSTAINABILITY CENTER was established in 2018 by Elaine Weidman Grunewald and Anna Felländer. The AI Center's vision is that a different and fairer approach to data, AI, and future technologies is possible to achieve. The AI Sustainability Center supports an approach in which the positive and negative impacts of AI on people and society are as important as the commercial benefits or efficiency gains. We call it Sustainable AI.

Contents

Recommendations		V
Abstract		VI
Introduction: Sustainable AI		VII
Objective: Inventory of the state of knowledge of sustainable AI		VIII
Design of the study		VIII
I	**Review of AI and ETHICS**	**10**
	1. Bias	12
	2. Accountability	16
	3. Abuse and malicious use	18
	4. Transparency and explainability	20
II	**Bibliometric literature review**	**24**
	Methodology	24
	The literature on sustainable AI	26
	Areas of research	27
	Summary of the bibliometric analysis	32
III	**In-depth studies**	**33**
	Medicine: Future health care challenges	33
	Telecom	37
	Digital platforms	39
Appendix 1		42

Recommendations

1. THERE ARE MANY REGULATORY ISSUES CONCERNING ARTIFICIAL INTELLIGENCE (AI). A more focused approach to these issues is urgently needed. This applies both to ethical frameworks as well as interpretations of current regulations in relation to newly evolved practices and methodologies. Regulatory authorities also need to be encouraged and educated to keep up with development of the technologies and methods for societally applied AI.

2. MULTIDISCIPLINARY AND INTERDISCIPLINARY RESEARCH ON APPLIED AI IS NEEDED to gain a greater understanding of the challenges posed by the technologies. This includes, among other things, bias issues, accountability, and the degree of transparency that is desired depending on context or application. AI's complex structure and its implications for society demand in-depth knowledge from different scientific disciplines such as engineering science, social sciences, medical science and the humanities. Research on sustainable AI requires collaborative efforts not only within academia but also between academia, business sectors and the public sector.

3. TRUST IS ESSENTIAL if we are to fulfill the promise and value that AI can bring in sectors such as retail, finance, health care, and more. It is crucial to improve knowledge and understanding of social bias and the relationship between explainability/transparency and accountability with regards to trust and social acceptance of AI.

Abstract

ARTIFICIAL INTELLIGENCE (AI) and rapid developments in machine-learning carry huge potential benefits. But whether these values will be realised in a sustainable manner is yet unknown. This report assesses that ethical, social and legal aspects have not been sufficiently incorporated and tested in research studies, or in the design and implementation of AI systems. This leads to unintended, negative consequences and risks involved in the implementation of AI in society.

We have focused on four problematic areas: 1. Bias; 2. Accountability; 3. Abuse and malicious use; 4. Transparency and explainability. By conducting an inventory of the state of knowledge of ethical, social, and legal challenges related to AI and machine-learning, this report identifies the areas of knowledge that require further study. Our conclusion is that there is a need for a multidisciplinary and interdisciplinary approach to research in this area, to enable the potential benefits of AI to develop in a sustainable manner. To do this, the report includes 1. A broad review of reports and studies that focus on ethical and sustainable AI; 2. A quantitative and bibliometric analysis of published materials in the combined fields of AI and ethics; and 3. In-depth studies of health and social care issues, telecom and digital platforms.

This knowledge review is a part of the Swedish Vinnova-funded project *"Hållbar AI – AI Ethics and Sustainability"*, which, among other things, is intended to gather a multidisciplinary consortium of relevant actors from academia and the business sector in order to identify unintended, negative consequences of AI.

Introduction: Sustainable AI

THE PRESENT KNOWLEDGE review was conducted within the Swedish Vinnova-funded project Hållbar AI – AI Ethics and Sustainability, led by Anna Felländer. The project is a part of a programme for challenge-driven innovations (UDI), and constitutes the start-up project of Stage 1. This project focuses on a key challenge: ethical, social, socio-economic and legal aspects have not been adequately integrated and tested in research, design and implementation of AI systems. The risk is that the implementation of AI and machine-learning applications in society could lead to unintended, negative, ethical and socio-economic consequences, e.g., in relation to consumer markets. In order to address this challenge, our vision is to establish a level of interdisciplinary competence and provide tools that enable organisations to meet certain standards and eventually receive certification. In this way, AI's potential could develop more sustainably.

Initially, the project convened a consortium of relevant actors from academia and the business sector in order to identify unintended, negative consequences of AI. For example, long-established prejudices may be reinforced by bias, thereby leading to unintended consequences. Other issues are, for example, whether AI applications are programmed to learn at a sufficient rate as well as inadequate knowledge and understanding of the impact of algorithms on continually evolving data. Furthermore, ethical evaluations are sometimes left to the discretion of the individual/individuals tasked with designing the algorithms. In Stages 2 and 3, the Center will increase its competency and develop testbeds, pilot projects and other activities. One of the Center's goals is to establish standards or a certification of ethical governance and management of data and AI for organisations and regulatory authorities. The first stage of the project also deals with inventorying the state of current knowledge, initiatives and practical examples, both in Sweden as well as internationally.

Objective: Inventory of the state of knowledge of sustainable AI

THE OBJECTIVE OF this report is to conduct an inventory of the state of knowledge in the areas of ethical, social and legal challenges pertaining to artificial intelligence and machine-learning. This is based on the fact that a need has been identified to develop an innovative approach that offers methods to address challenges and establish what areas of knowledge require further study.

Design of the study

AT PRESENT, OUR knowledge of the ethical, social and legal consequences of AI is fragmented; this applies both to scientific disciplines as well as how such knowledge is published and disseminated. Since we are dealing here with a new field of knowledge, it can be concluded that a sort of conceptual development is under progress in this area, often expressed in the form of reports and white papers rather than peer-reviewed journals and conferences. As a result, an inventory of this area requires a relatively broad approach. We have divided the types of published materials into three main categories:

- Reports, policies and conceptual work
- Peer-reviewed articles in general
- Bibliometric reviews of literature retrieved from Web of Science (WoS)

SECTION 1 CONSISTS of a broad review of the large amount of reports and studies published in the areas of ethics and sustainable AI. These can be found in reviewed scientific journals and conference minutes, but also strikingly often in reports by expert groups, research institutes and government agencies. The latter category is an indication of how strongly these practically anchored issues have developed in the last 3 to 5 years, and is furthermore backed up by the bibliometric analysis presented in SECTION 2. SECTION 1 is divided, albeit far from exhaustively, into four key categories:

- Bias
- Accountability
- Abuse and malicious use; and
- Transparency and explainability.

SECTION 2 consists of a quantitative and bibliometric analysis of published materials retrieved from the combined areas of AI, broadly speaking, and ethics, broadly speaking. This section mainly addresses peer-reviewed studies written in English.

SECTION 3 delves deeper into three areas in order to establish some of the practical consequences of AI, both potential benefits as well as challenges, that various business sectors and other areas of focus face. We focus on the following three areas:

- Medicine: Health and social care;
- Telecom; and
- Digital platforms.

The latter category is, perhaps, less established than the previous two, but in this section, we apply findings from the social sciences regarding the social relevance of data-driven organisational logic, as employed in digital platforms. Given the economies of scale of these platforms and our daily use of these services, how AI is implemented is crucial when it comes to moderating and governing how these platforms are used.

I
Review of AI and ETHICS

MANY COMPANIES ARE actively addressing the challenges that are the subject of this report. Some companies have also voiced their position and opinions on the possibilities of preventing unintended, negative consequences resulting from systems and technology based on artificial intelligence (AI). This includes tech companies such as IBM[1], Microsoft[2] and Google[3]. The European Union (EU) has also initiated research projects[4] and has published reports[5][6] that underline the importance of defining policies to address the ethical challenges in the wake of autonomous systems and AI. Furthermore, a number of influential industrial organisations, standards organisations, and research institutes are also actively addressing the issue, for example, IEEE[7], ITU[8], ACM[9], ANE[10] and AI Now[11]. All of

1 https://www.ibm.com/blogs/policy/trust-principles
2 https://www.microsoft.com/en-us/ai/our-approach-to-ai
3 https://www.blog.google/technology/ai/ai-principles/
4 http://europa.eu/rapid/press-release_IP-18-3362_en.htm
5 http://publications.jrc.ec.europa.eu/repository/bitstream/JRC113826/ai-flagship-report_online.pdf
6 http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf
7 https://standards.ieee.org/industry-connections/ec/autonomous-systems.html
8 https://www.itu.int/en/ITU-T/AI/Pages/ai-repository.aspx
9 https://www.acm.org/code -of-ethics
10 https://ipaper.ipapercms.dk/IDA/ane/report/#/
11 https://ainowinstitute.org

these research projects use factual examples to highlight the need for further research in this area. In some cases, they also offer recommendations for technological developments and their applications, in order to minimize the risk of negative consequences, where possible. However, no general standards or commonly shared guidelines have been adopted, as yet.

Additionally, the IEEE (the Institute of Electrical and Electronics Engineers) has initiated a program to establish a certification system of ethical approaches to autonomous and intelligent systems (ECPAIS[12]). The purpose of the program is to focus on areas such as transparency, accountability, and to minimize algorithmic bias. There are also examples of conceptual approaches toward "AI for good",[13] and calls for action, such as *the Montreal Declaration for a responsible development of artificial intelligence*.[14]

The EU Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) has drafted a list of ethical principles (published on December 18, 2018). This draft has been circulated for comments and the ethics guidelines for trustworthy AI was published in April 2019.[15] One of the participants in our project group, Fredrik Heintz, is a member of this European expert panel.

With regards to data-driven and algorithm-driven systems and potential consequences of applied AI, there is a growing understanding in the literature that legitimacy, accountability, and transparency are of crucial importance. A relatively new field has emerged that focuses on Fairness, Accountability and Transparency, or FAT, for short. FAT highlights the fact that algorithmic systems are used in a number of different situations where vast amounts of data (Big Data) are employed in order to screen, categorise, rate, recommend, "personalize" and in other ways

12 https://standards.ieee.org/industry-connections/ecpais.html

13 AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations https://link.springer.com/article/10.1007/s11023-018-9482-5

14 Montreal declaration responsible AI. https://www.montrealdeclaration-responsibleai.com/the-declaration?fbclid=IwAR0CjNOAlx0flYgpAqxq2gy05xGPka7YooNlhTjTEe5-qGMNRtL0oBN9EOo

15 https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

shape human experiences and relationships. Although there are benefits to many of these systems, there are also inherent risks, such as the codification and reinforcement of social bias, reduced responsibility, and increased information asymmetry between data producers (customers) and data holders.[16]

The following passage focuses on four categories that we assess as being key components when addressing the challenges posed to sustainable AI: bias, accountability, malicious use, and transparency.

1. Bias

A NUMBER OF cases have received attention for promoting unintended social bias, which is then reproduced or automatically reinforced by AI systems; often, in-depth studies are required for them to be noticed at all. Due to the complexities related to transparency issues, discovering the presence of reproduced, and even reinforced, social bias is a tricky task, and therefore we shall revisit this issue later in the text.

Some research groups have studied and discovered automated ad-distribution tools that contained gender biases that were more likely to distribute well-paid job ads to men than women.[17] Other studies conclude that popular image databases also have a gender bias, and regularly portray women performing kitchen chores while men are out hunting, resulting in a self-learning application that not only reproduced gender bias, but also amplified it.[18] In a widely criticized case of algorithm-assisted decision-making in the USA by public bodies based on recidivism prognoses,

16 For an in-depth socio-legal and legal scientific analysis of FAT, please see Larsson, S. (2019) "Artificiell intelligens som normativ samhällsutmaning: partiskhet, ansvar och transparens".

17 Datta, A., Tschantz, M.C., Datta, A. (2015). Automated Experiments on Ad Privacy Settings – A Tale of Opacity, Choice, and Discrimination. Proceedings on Privacy Enhancing Technologies. 1: 92–112, DOI: 10.1515/popets-2015-0007.

18 Published by Wired Magazine, 21 August 2017: https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/ ; Please see the study by: Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457.

i.e., the tendency of a convicted criminal to reoffend, the investigative journalism organisation ProPublica showed that the COMPAS system was more likely to incorrectly predict that black defendants represent a high-risk group, while simultaneously, and incorrectly, predicting the opposite in the case of white defendants.[19] Similar examples can be found in programmer Cathy O'Neil's much debated book *Weapons of math destruction: How big data increases inequality and threatens democracy*.

A scientific review of the three commercial, gender-determining image recognition systems show that the group that is most likely to be categorized incorrectly consists of women with darker skin.[20] This means, among other things, that services and applications based on these systems poorly serve groups of a certain physical appearance. The margin of error is significantly narrower for light-skinned men. In line with this, it has been observed that one of the most popular image databases, ImageNet, which contains around 14 million annotated images, largely contains images collected from a handful of countries, such as the USA and the UK. This has consequences for machine-learning with regards to cultural expressions; for example, searches for wedding gowns produce the standard white version commonly used in the USA, while Indian wedding gowns are categorised as "performance art" or "costumes".[21] When applications are programmed with this kind of bias, it can lead to situations such as cameras that automatically warn the photographer that the subject of the photograph has his/her eyes closed, based on stereotypical, masculine and light-skinned appearances. For example, the camera may

19 For an in-depth analysis, please see Caplan, R., Donovan, J., Hanson, L. and Matthews, J. (2018). Algorithmic Accountability: A Primer, NYC: Data & Society; and Larsson, S. (2019) "Artificiell intelligens som normativ samhällsutmaning: partiskhet, ansvar och transparens". The study was carried out and published by the civil rights-driven, investigative journalism organisation ProPublica (23 May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

20 Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (pp. 77-91).

21 Zou, J. & Schiebinger, L. (18 juli 2018). "AI can be sexist and racist — it's time to make it fair", Nature, comment. https://www.nature.com/articles/d41586-018-05707-8

determine that Asians always are blinking.[22]

The fact that search engines, which are largely automated and contain self-learning – i.e. artificially intelligent – elements, interact, reproduce and are to some extent a product of social, historical and cultural structures, was recently, and emphatically, demonstrated by American communications researcher Safiya Noble (2018). In her book *Algorithms of Oppression*: *How search engines reinforce racism* she presents a critical analysis that she refers to as "technological redlining", showing that data analyses covertly and structurally may discriminate against certain groups, and are often only discovered upon extensive scrutiny, after the event has occurred. One challenge here would seem to be that the relationship between inherent social structures and historically-based inequality is manifest in the data used to train self-learning algorithms. If the data contains social bias, this will be reproduced in its outcome.

Job ads are sometimes described as a particularly problematic area, with regards to bias.[23] This issue was raised anew in October 2018 in connection with Amazon's development of a self-learning tool used to judge work-seekers that was found to contain significant bias in favour of men, and awarded them top ranking.[24] This system had learned to prioritise job applications that to a great extent emphasised male characteristics, and downgrade applications from universities with a strong female presence. This example showed the unintended consequences of machine-learning applications, where the applied training materials unwittingly lead to unintended, and biased consequences.

There are examples of innovations that have been produced to counteract bias, e.g., the New York-based company Pymetrics, which offers what they call "neuroscience games and bias-free AI to predictively match people with jobs where they'll perform at the highest levels". Their method

22 Zou, J. & Schiebinger, L. (18 juli 2018).

23 https://www.forbes.com/sites/tomaspremuzic/2018/05/27/four-unethical-uses-of-ai-in-recruitment/amp/

24 https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

of job matching enables the job recruiter to allow a selection of high-performing employees to play Pymetrics' games, which have been designed to assess characteristics such as memory, emotional recognition, risk propensity, sense of justice, and ability to stay focused.[25] Pymetrics determines what characteristics can be linked to high-performance in relation to specific company positions. Jobseekers then go through the same process and are assessed using artificial intelligence rather than humans in order to avoid any bias that could arise from the jobseeker's name, gender, skin colour, ethnicity, age and CV. Finally, Pymetrics recommends that the job recruiter employs jobseekers that display characteristics that are similar to their top employees – thereby assessing their "inner" rather than their "outer" qualities. However, there are likely challenges associated with this method with regards to desired characteristics and unintended effects, but it does show that there are alternative methods.

To a certain extent, systematic bias may arise not only as a result of the data used to train systems, but also as a result of value-based preferences held by system developers and users of the system. For example, the "legacies of bias" is discussed in an AI Now report stating that AI is not impartial or neutral: "Technologies are as much products of the context in which they are created as they are potential agents for change."[26] Our understanding and experiences of our surroundings are based on previous experiences, perceptions, and how we envision future goals. Cognitive science, for example, is a broad area of research that in recent years has begun to conduct studies[27] of how our perception governs our interactions and our interpretations of results produced by AI and systems based on self-learning machines.

25 https://www.pymetrics.com/about/

26 Campolo, A., Sanfilippo, M., Whittaker, M. & Crawford, K. (2017) AI Now 2017 Report. AI Now Institute at New York University, p. 18.

27 Kliegr, T., Bahník, Š., & Fürnkranz, J. (2018). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. arXiv preprint arXiv:1804.02969. https://arxiv.org/pdf/1804.02969.pdf

2. Accountability

ACCOUNTABILITY ISSUES WITH regards to limitations and unintended consequences of AI applications in autonomous systems is increasingly becoming a hotly debated topic in news media, and a growing body of literature has begun to address concepts such as *algorithmic accountability* and responsible AI. Algorithmic accountability, according to Caplan's et al. report, published in Data & Society,[28] deals with the delegation of responsibility for damages incurred as a result of algorithmically-based decisions producing discriminatory or unfair consequences.[29] This can also be applied to accountability issues in developments in algorithms and their social effects and consequences. In the event of damages incurred, responsible systems should include a mechanism for redress.

Legal scientists such as Hildebrandt[30] have raised the issue of the "agency of things", i.e., the fact that AI allows a greater degree of perpetually self-learning autonomy, as well as the link between autonomy and fairness. Additionally, as socio-legal researcher Larsson concludes, issues inevitably emerge in connection with the *agency of things*, or the *agency* of software processes when they become endowed with the ability to survey and learn from vast amounts of information, not least in the context of automated decision-making processes.[31]

One area with regards to accountability issues is the introduction of self-driving vehicles. In the event of an accident, who should be held accountable? Autonomy, which, in the case of data-driven applications is very much dependent on algorithms designed to perform necessary functions, is a key area of focus with regards to self-driving vehicles, but it also raises issues of accountability. Regulations for self-driving vehicle

28 Caplan, R., Donovan, J., Hanson, L. and Matthews, J. (2018). Algorithmic Accountability: A Primer, NYC: Data & Society.

29 Jfr. Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. Digital Journalism, 3(3), 398-415;

30 Hildebrandt, M. (2015). Smart Technologies and the Ends of Law, UK & USA: Edward Elgar Publishing; see also Larsson (2019).

31 Larsson 2019, s. 351.

technology are currently being drafted in a number of countries, including Sweden (SOU 2018:16),[32] where accountability is of crucial importance in traffic accidents – a topic that has been discussed for some time in the literature.[33] These problems have been highlighted not least in connection with deadly accidents involving autonomous vehicles, resulting in a need to evaluate and judge this mix of software, (safety) drivers, vehicle hardware, and external events. In 2016, a Tesla Model S equipped with radar and cameras determined that a nearby lorry was in fact the sky, which resulted in a fatal accident. In March 2018, a car used by Uber in self-driving vehicle trials hit and killed a woman in Arizona, USA, which raised extensive discussions on responsibility issues and self-driving vehicles in public traffic. Even if comparisons between traffic situations with and without self-driving vehicles were to show that autonomous vehicles are significantly safer, incidents like this will continue to have a detrimental impact on people's trust and their acceptance of highly autonomous vehicles.

There are articles that address this issue and propose some form of global, international authority to create the necessary regulatory framework (laws, policies).[34] Other researchers have compared accountability issues in the context of AI systems to the healthcare sector with regards to medicines[35]. A more dynamically regulated system, such as the aforementioned, requires that algorithms be tested in live situations to ensure that any potential "side effects" are minimised as far as possible.

32 On 1 July 2017, the government enacted new rules for self-driving vehicles that made it easier to test self-driving vehicles in public traffic (Regulation 2017:309, please see SOU 2016:28). The regulation also provides rules that require a human driver be present either in or outside the vehicle. On 7 March 2018, a final report on self-driving vehicles was submitted to the government (SOU 2018:16; please see Dir. 2015:114) in which the division of responsibility and data protection represent a significant portion of the report.

33 cf. Hevelke A., & Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. Science and Engineering Ethics 21(3): 619–630.

34 Erdelyi, Olivia Johanna and Goldsmith, Judy, Regulating Artificial Intelligence: Proposal for a Global Solution (February 2, 2018). 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2–3, 2018, New Orleans, LA, USA doi/10.1145/3278721.3278731. Available at SSRN: https://ssrn.com/abstract=3263992 http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf

35 London, A. J., & Danks, D. Regulating Autonomous Vehicles: A Policy Proposal. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_111.pdf

3. Abuse and malicious use

Many researchers argue that some accountability for abuse and malicious use of AI should lie with the designers and developers of AI software.[36] Autonomous weapons and Max Tegmark's et al. initiative to take a *Lethal Autonomous Weapons Pledge* can be mentioned.[37] There is, however, a less dramatic threat scenario, as described by Brundage et al., which does not necessarily, or explicitly, pertain to militarisation. For example, advanced forms of cyber-attacks such as automated hacking, or remote control of online, autonomous vehicles to attack people, e.g., by steering the vehicle into crowds. This also includes political and polarising activities that employ botnets to influence elections,[38] or to create division on various matters, as can be seen in the ongoing "anti-vaxx" discussions in the USA.[39] The research group focusing on the malicious use of AI therefore calls for AI developers to promote a stronger culture of responsibility with regards to how their tools can be used, which emphasizes the need for education, ethical standards and norms.[40]

Another challenge that needs to be addressed is the fact that self-learning software may expose inherent social bias and partiality, and that the software design, in itself, may become normative. The problem, then, has to do with the question of accountability, both with regards to how the tools may be used as well as the values that autonomous design actually expresses and reproduces. This issue has been addressed in relation to digital platforms,[41] search engines and social media which may not only reproduce discrimination, racism and inequality, but, in fact, may also strengthen these structures.

36 Brundage, M. et al. (2018) The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. https://maliciousaireport.com.

37 https://futureoflife.org/lethal-autonomous-weapons-pledge/

38 Bastos, M.T., & Mercea, D.(2017). The brexit botnet and user-generated hyperpartisan news. Social Science Computer Review, https://doi.org/10.1177/0894439317 734157.

39 e.g. Broniatowski, D.A. et al. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate, American Journal of Public Health, published online before print. DOI: 10.2105/AJPH.2018.304567

40 Brundage et al., 2018, s. 7.

41 Larsson, S. (2018) "Sju nyanser av transparens: Om artificiell intelligens och ansvaret för digitala plattformars samhällspåverkan," Andersson Schwarz, J. & Larsson, S. (ed.) Plattformssamhället. Den digitala utvecklingens politik, innovation och reglering. Stockholm: Fores.

Some studies have pointed out that there are a range of software products that could be used to burst so-called filter bubbles, while simultaneously pointing out that the concept of democracy is an ambiguous one and covers a number of different interpretations. We can conclude that software developers also need to be made aware of this issue.[42]

The developers who design these systems should arguably bear some accountability for ensuring that these systems do not perpetuate unintended, built-in bias. How to go about eliminating bias and determining who is accountable is currently being debated in many fields. There are well-proven and documented methods for testing algorithms for partiality and social biases[43]. However, the American CFAA Act (the Computer Fraud and Abuse Act) allows companies to block any such tests of their products. One study[44] addresses the fact that the CFAA Act, in fact, violates US law (a number of lawsuits have been filed in the USA as a result).

One problem is that effective methods for testing algorithms for bias often require several, alternative (fake) profiles (users) in order to identify different outcomes that can then be linked to different kinds of user profiles (i.e., gender, age, place of residence, etc.). Another method involves repeatedly sending the same requests to the system to see whether the outcomes differ in anyway (so-called "scraping"). However, the CFAA Act forbids both methods based in the argument that they use the system in an "incorrect" manner. However, this also may prevent conducting meaningful tests for potential bias in algorithms, which would serve the public interest and, in the case of the USA, probably violates civil rights, as laid out in the United States Constitution.

42 Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: democracy and design. Ethics and Information Technology, 17(4), 249-265.

43 Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry, 1-23. http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf

44 Patel, K. S. (2018). Testing the Limits of the First Amendment: How Online Civil Rights Testing is Protected Speech Activity. Columbia Law Review, 118(5). https://columbialawreview.org/content/testing-the-limits-of-the-first-amendment-how-online-civil-rights-testing-is-protected-speech-activity/

4. Transparency and explainability

IN THE COMPUTER science literature, the area of interpretable and explainable machine-learning, sometimes abbreviated as XAI, has been an object of research for some time; but a critical review reveals a need to more clearly define the issue,[45] not least in relation to the growing application of machine-learning,[46] and that disciplines such as social psychology and cognitive science could make important contributions.[47] A well-known problem with regards to accountability in relation to algorithm-driven processes is the lack of transparency, sometimes referred to as black box systems.[48] Much of the issues surrounding accountability are related to how we perceive and understand the events in focus, which highlights the importance of developing our understanding of the relationship between transparency and socially and commercially applied AI, although transparency should not be seen as a one-size-fits-all solution.[49]

In 2018, The EU Commission initiated a study, to be concluded in 2019, that aims to analyse so-called algorithmic transparency in order to increase awareness and establish a sound knowledge base for dealing with the challenges and potential benefits of algorithmically-assisted decision-making:

> Algorithmic transparency has emerged as an important safeguard for accountability and fairness in decision-making and for opening to scrutiny the way access to information is mediated online, especially on online platforms.[50]

45 Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 51(5): 93.

46 Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In IJCAI-17 Workshop on Explainable AI (XAI).

47 Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. Vol 267: 1-38. https://doi.org/10.1016/j.artint.2018.07.007.

48 Cf. Pasquale, F. (2015). The Black Box Society. The Secret Algorithms That Control Money and Information, Harvard University Press;

49 Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973-989.

50 The EU Commission (April 25, 2018) Algorithmic Awareness-Building. https://ec.europa.eu/digital-single-market/en/algorithmic-awareness-building

However, when calling for increased transparency, it is important to nuance our understanding of transparency: for whom is the transparency aimed? How is it conveyed? And for what purpose? As mentioned, transparency does not solve every problem. There are conflicting interests with regards to transparency and situations where full transparency can, in fact, impede fairness and attempts to avoid bias. Larsson lists seven challenges to transparency in relation to AI and machine-learning processes.[51] One societal challenge is to determine how to weigh conflicting interests. Points 1 and 2 below represent conflicting interests with regards to transparency, and 3-7 represent different kinds of challenges with regards to knowledge and transparency, including XAI. Several studies have also underlined the need for perspectives based in the social sciences, psychology or philosophy to complement interdisciplinary research on AI and explainability.[52]

1. **PROPRIETORSHIP:** software and data are proprietary works, (and thereby incompatible with transparency); i.e., it may not be in a company's best interest to divulge how they address a particular problem, as may be the case when a product is commercialised and scaled up for commercial purposes. Many companies view their software and algorithms as valuable "recipes", trade secrets that are absolutely key to maintaining their position in a competitive market.[53]
2. **PREVENTING ABUSE** ("gaming"): transparency can be abused to counteract the intended objective and enable abuse or manipulation to gain advantages, e.g. Twitter's trending function, or when distributing social benefits, or other processes that involve profiling or rating systems.[54]

51 Larsson, S. (2019). "Artificiell intelligens som normativ samhällsutmaning: partiskhet, ansvar och transparens" i Banakar, Dahlstrand & Ryberg-Welander (ed.) Festskrift till Håkan Hydén. Lund: Juristförlaget; Larsson, S. (2018) "Sju nyanser av transparens: Om artificiell intelligens och ansvaret för digitala plattformars samhällspåverkan," i Andersson Schwarz, J. & Larsson, S. (ed.) Plattformssamhället. Den digitala utvecklingens politik, innovation och reglering. Stockholm: Fores.

52 e.g. Mittelstadt, B., Russell, C., & Wachter, S. (2018). Explaining explanations in AI. arXiv preprint arXiv:1811.01439.

53 E.g. Spiekermann, S., & Korunovska, J. (2016). Towards a value theory for personal data. Journal of Information Technology, 23(1): 62-84. doi:10.1057/jit.2016.4.

54 e.g. Caplan et al. (2018).

3. **COMPETENCE AND LITERACY:** the ability to understand and assess algorithms, how they are applied to data, and their consequences in everyday situations requires competence, sometimes referred to as data literacy or algorithmic literacy.[55]
4. **CONCEPTS, METAPHORS AND TERMINOLOGIES:** the language, metaphors and symbols used to explain AI processes have a direct impact on how we conceptualize and understand such explanations, which, in turn, is related to acceptance and trust.[56]
5. **MARKET COMPLEXITY:** a combination of proprietary arrangements and data-driven markets that can be seen as complex "ecosystems" in which data is brokered and transferred to a number of actors. This also includes the often commercially motivated practice of using trackers, such as third-party cookies and pixels, meaning that it becomes difficult to track where the data travels.[57]
6. **DISTRIBUTED, PERSONALISED OUTCOMES:** the outcome of consumer-profiling services that attempt to "personalise" their services, their prices or marketing campaigns – and pose a challenge not least to regulatory oversight.[58]
7. **ALGORITHMIC COMPLEXITY:** self-learning algorithms are endowed with a level of independent autonomy that prevents actual oversight of how algorithms solve problems – a human viewer may only able

55 For more on algorithms, please see Haider & Sundin (2019) "Algoritmernas roll i plattformssamhället. Vad är algoritmer, och vad gör dem till så viktiga komponenter i plattformssamhället?" i Andersson Schwarz, J. & Larsson, S. (ed.) Plattformssamhället. Den digitala utvecklingens politik, innovation och reglering. Stockholm: Fores; Haider & Sundin (2019) Invisible Search and Online Search Engines: The ubiquity of search in everyday life. Chicago: Routledge Studies in Library and Information Science, which focuses on the importance of media and information literacy in relation to search engines.

56 Our understanding of abstract, e.g. digital, phenomena can have a decisive impact on how they are regulated as well as our normative understanding of them. For an extensive study of the legal implications of metaphors and conceptual metaphors in relation to digital phenomena, please see Larsson (2017) Conceptions in the Code. How Metaphors Explain Legal Challenges in Digital Times. Oxford University Press.

57 See Pasquale, F. (2015). The Black Box Society. The Secret Algorithms That Control Money and Information, Harvard University press. The complexities involved in the commercial setup are concisely explained in Christl, W. (2017). Corporate Surveillance in Everyday Life: How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions. Vienna: Cracked Labs.

58 This is further analysed in the context of consumer protection in Larsson, S. (2018a). Algorithmic Governance and the Need for Consumer Empowerment in Data-driven Markets, Internet Policy Review 7(2):1–12.

> to see whether the problem has been solved or not. This may result in a higher likelihood of a certain outcome, which, in practice, could be applied to increase profitability and sales, or to improve accuracy in diagnosis, but may not necessarily describe in detail how these outcomes were achieved.[59]

Several studies have confirmed this particular point, and researchers underline the need for "auditability",[60] i.e. to allow third parties to scrutinise and study how an algorithm performs, such as research projects that study discriminatory practices of digital platforms.[61] This has also been described as an important component of the EU Commission Expert Panel's draft of ethical guidelines to ensure reliability in AI systems, thereby making it possible to track earlier decisions that led to certain consequences.[62]

There are also ongoing, extensive discussions on automated decision-making and the GDPR data protection regulation, and whether individuals are "entitled to an explanation" with regards to decision-making based on profiling.[63] This endeavour is seen in some circles as a suitable mechanism for accountability and transparency in connection with automated decision-making. How to develop explainability in relation to automated decision-making based on personal data analyses and profiling is an issue that will become increasingly important in the future.

---

59 This is known as AI-explainability (XAI) in AI research, please see IEEE, 2018, cf. Wachter et al. (2017). Another aspect worthy of discussion in the context of self-learning models, are the differences between interpretability, explainability and comprehensibility, please see Guidotti, r., Monreale, A., ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys (CSUr), 51(5): 93.

60 Diakopoulos, N., & Friedler, S. (2016). How to hold algorithms accountable. MIT Technology Review, 17(11);

61 Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry, 1-23.

62 The European Commission's High-level Expert Group on Artificial Intelligence (18 December 2018) DRAFT ETHICS GUIDELINES FOR TRUSTWORTHY AI. Working Document for stakeholders' consultation Brussels.

63 Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 7(2), 76-99.

II
Bibliometric literature review

ONE WAY OF gaining an overview of a corpus is by employing bibliometric analyses; i.e., statistically analysing texts and corpora as well as their descriptive meta-data by analysing links and co-citations in journals, among other things. This bibliometric review was conducted in November 2018 in collaboration with bibliometrician Fredrik Åström, Lund University.

The objective of the bibliometric analysis is to describe, using quantitative analyses of the literature in the field, the most salient aspects of AI research. These aspects cover fields of research that study AI issues by analysing journals cited in AI studies. The contents of the papers are analysed by studying the concepts and terminology used in the articles.

Methodology

IN ORDER TO identify research literature with a focus on AI issues, we used the Web of Science databases (WoS), which mainly indexes articles published in international scientific journals. Searches were conducted

using search strings that were based on a combination of relevant terms that were then matched against headings and titles, abstracts and keywords. Please note the combined results of literature that pertains to AI and machine-learning, and issues dealing with ethics, accountability and social bias; i.e., topics that are of crucial importance to *sustainable AI*.

> ("artificial intelligence" OR "machine learning" OR "deep learning" OR "autonomous systems" OR "pattern recognition" OR "image recognition" OR "natural language processing" OR "robotics" OR "image analytics" OR "big data" OR "data mining" OR "computer vision" OR "predictive analytics")

AND

> ("ethic*" OR "moral*" OR "normative" OR "legal*" OR "machine bias" OR "algorithmic governance" OR "social norm*" OR "accountability" OR "social bias")

Furthermore, we limited our searches to the terms, "Article", "Book Chapter", "Letter", "Proceedings Paper" and "Review". This process yielded 2,706 published articles, and their related meta-data was downloaded from the WoS database and analysed using Bibexcel[64] and VOSviewer[65].

To discover which areas of research are involved in AI research, we used the "journal co-citation analysis" method,[66] using Bibexcel to excerpt reference lists from the articles included in the analysis; this allowed us to locate the journals in which the articles were published. The results were then analysed by studying the frequency of co-cited journals in the reference lists. Based on the frequency of co-cited journals, a network of journals begins to emerge which can then be visualised using VOSviewer. VOSviewer reads frequency of co-citations and plots them according to

64 https://homepage.univie.ac.at/juan.gorraiz/bibexcel/

65 http://www.vosviewer.com/

66 McCain, K.W. (1991). Mapping economics through the journal literature: An experiment in journal cocitation analysis. Journal of the American Society for Information Science, 42(4):290-296.

data proximity, where journals that are often co-cited are gathered close to each other while less cited journals are placed further apart. Based on this, clusters of frequently co-cited journals are produced and used to represent different areas of research.

The contents of the studies are similarly mapped by studying concurrent terms used in the literature, so-called "co-word analysis" (Callon et.al, 1983).[67] Instead of retrieving cited journals, this allows us to excerpt terminology used in the articles' titles and headings, abstracts, and keywords used to describe the articles. To avoid irrelevant terms, we began by conducting a relevance analysis, and then analysed the relevant terms based on frequency of co-citation.

The literature on sustainable AI

WE BEGIN WITH a descriptive analysis of the AI literature (AI *and* ethics) and its development over time (Figure 1). The literature within this sample of AI research was largely produced in 2010, and 75% of it was published later than 2011. The search terms used to define AI research and literature also located the occasional, odd article published between 1970 and the early 1990s. Between 1996 and 2011, we see an annual increase in published articles, beginning with only a dozen or so up to almost a hundred; and between 2012 – 2016, we see a steep increase in published articles, which then increases by roughly 100% every other year.

67 Callon, M., Courtial, J-P., Turner, W.A., Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. Social Science Information, 22(2), 191-235.

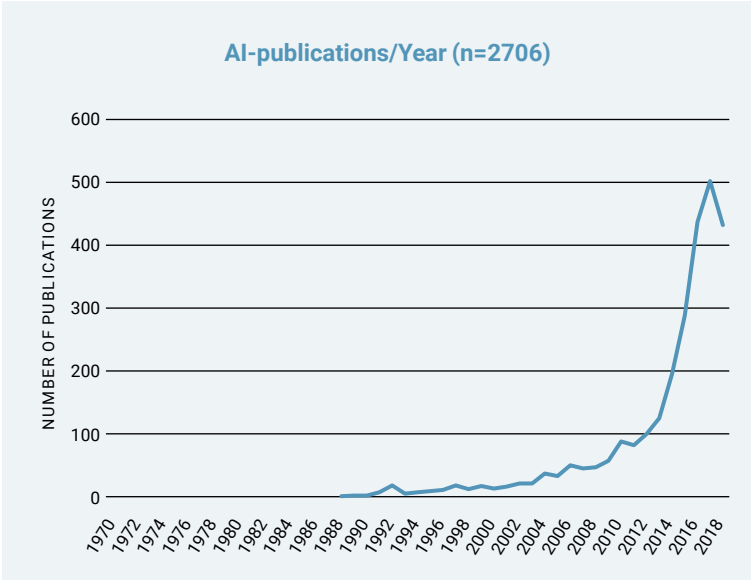


FIGURE 1. Published articles per year: sustainable AI.

The sudden drop in articles in 2018 is due to the fact that our research was completed before the year ended and so were were not able to access all AI articles for the year. However, an additional 300 articles were published between May 2018 – when the first keywords trial was run – and the end of November, which is when the final study was conducted.

Areas of research

BASED ON THE 2,706 published articles, and the 141,083 references listed in these articles, the 731 articles that were cited at least 10 times have been analysed to examine how often they appear in the articles. The size of the nodes, i.e., cited journals, represents how often each journal has been cited respectively, while the distance between the journals, and the links between the nodes, represent how frequently the journals have been co-cited.

The colour coding is based on a statistical cluster analysis of frequency of co-cited journals in which the clusters represent different areas of research (Figure 2). Based on these clusters of cited journals, we have presumed that they also represent the fields of research in which the studies were conducted; e.g., in the case of clusters of cited journals related to psychology, we have presumed that researchers and articles that cite these journals can generally be linked to the field of psychology.

This analysis identifies five main clusters. In the middle of the chart, we see that "Science" and "Nature" are the most cited journals. We have not linked them to any specific area of research due to their multidisciplinary nature. In the top left, we see a cluster of legal journals, such as "Harvard Law Review" and "Stanford Law Review". In the middle of the upper half of the chart, we see two clusters, the yellow one representing journals related to psychology, e.g. "Psychological Review" and "Trends in Cognitive Science", and the blue one representing computer science journals such as, "Artificial Intelligence", "Lecture Notes in Computer Science" and "Machine Learning". In the top right, we see a large cluster of medical journals. Aside from journals that do not have a specific orientation, we also see journals that represent different kinds of medical fields such as "Neurology" and "Brain", "Neuroimage" and "Radiology", and "European Journal of Human Genetics".

The medical cluster contains a number of cited journals that deal with medical ethics and information management, e.g. "Journal of Law, Medicine and Ethics", "American Journal of Bioethics", "Journal of Medical Ethics", and "Journal of the American Medical Information Society". In the bottom left, we see a cluster of social science journals, the majority of which deal with informatics and communication sciences – e.g., "Communications of the ACM", "Information Systems Research", and "Information, Communication & Society" – but there are also journals that lean toward business economics such as "Harvard Business Review" and "Management Information Systems Quarterly". We also find journals that represent different fields within the social sciences, such as sociology and political science; but there are also journals such as "Science, Technology

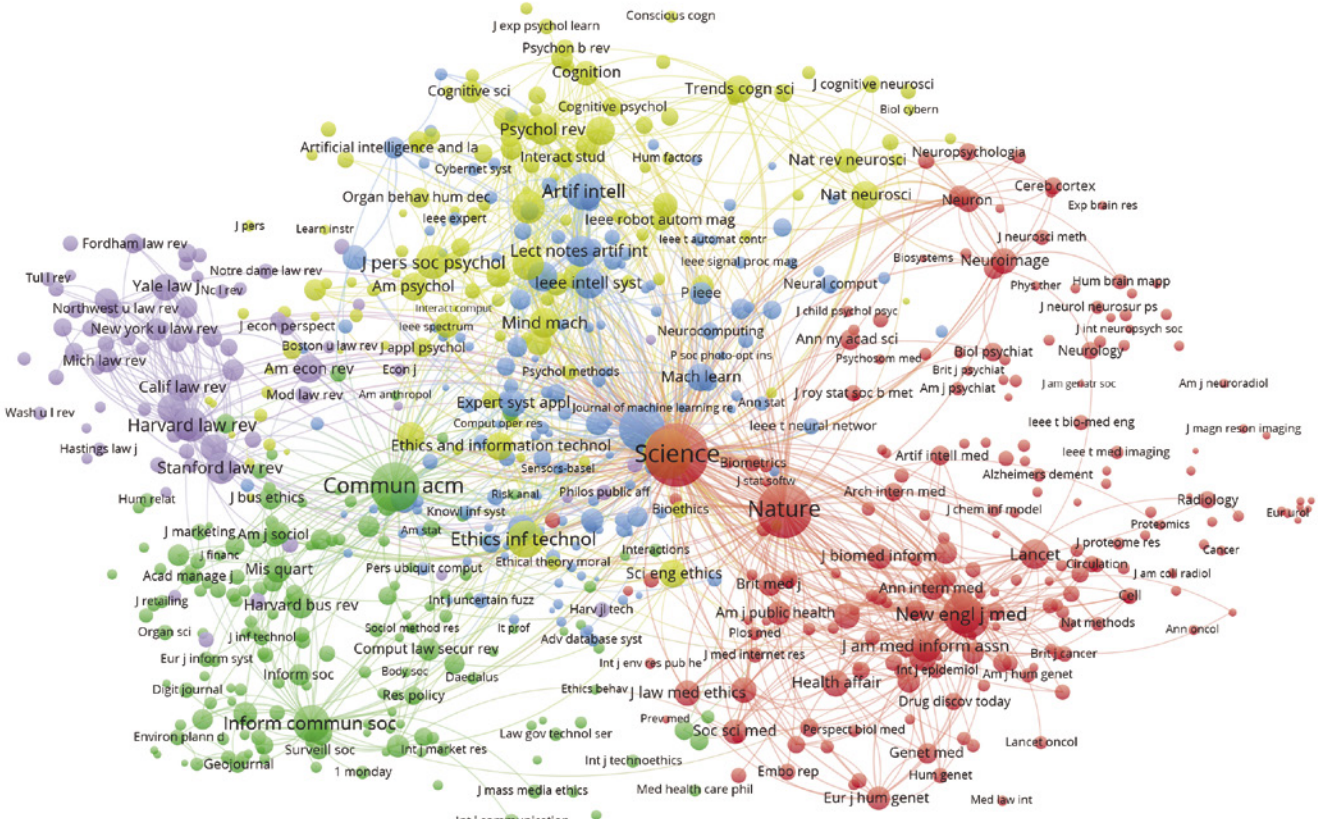


FIGURE 2. Network of co-cited journals (731 journals, cited 10 times or more).

and Human Values", "Social Studies of Science", and "Philosophy and Technology", that could be said to represent research fields that adopt a more critical approach to R&D, as well as a number of journals that address ethical perspectives on information management, technology and media.

To examine the contents of these studies, the terms used in the titles, headings and abstracts, and keywords that describe the articles were analysed. Based on a total of 40,349 terms, a relevance analysis was conducted to eliminate irrelevant terms and rank the most relevant, including a sample of terms that recur at least 20 times. These 306 terms were analysed by studying how often they appear together in the 2,706 published articles, using the same method as used for cited journals. However, instead of visualising networks, the links are visualised according to density, where larger groups of clustered terms are visualised partly by locating them close to each other and partly by grouping them in dark red fields to indicate dense groups of terms, and light red fields to indicate sparsely populated areas.

The above visualisation shows three main clusters of terms. In the upper half, we see terms that relate AI issues and developments in technology to ethical issues, represented by terms such as "robotics", "autonomous systems" and "engineering", and "ethics" (also in connection with, e.g., "robot" or "machine") and "morality". The bottom half of the chart shows terms that can be linked to developments in technology and data analysis, such as "machine-learning", "algorithm", "pattern recognition", "neural network" and "support vector machine". In the bottom left corner, we see a cluster of terms that mainly relate to issues concerning data security and privacy, e.g. "privacy", "data protection", "confidentiality", and "informed consent". These terms are also linked more specifically to issues within medical research and health research, which is reflected in the use of terms such as "health data", "disease", "clinical trials" and "treatment".

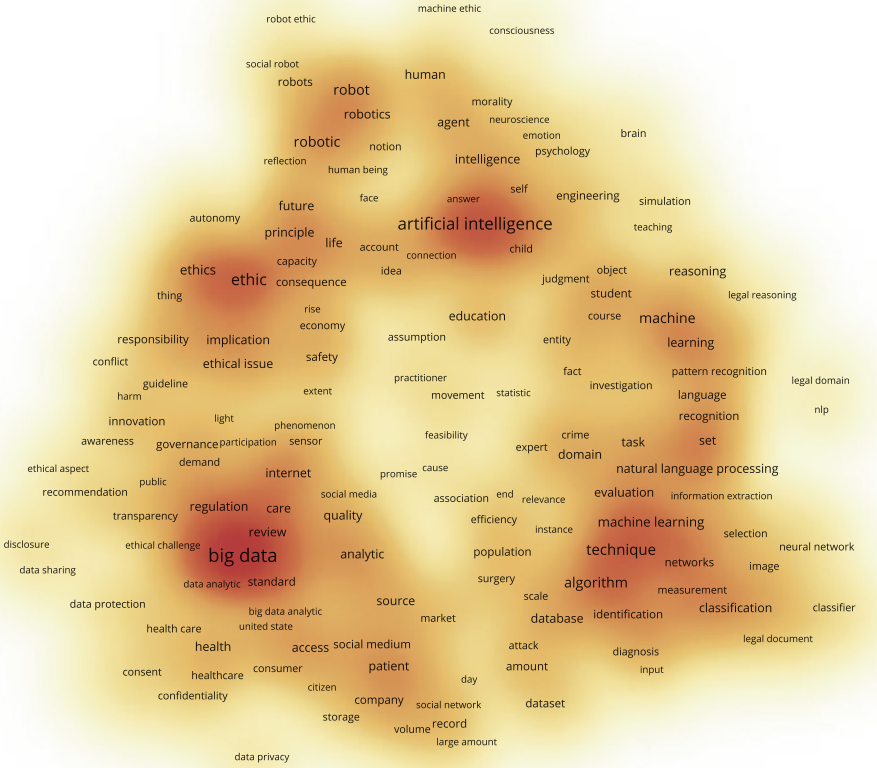


FIGURE 3. Concurrent terms (306 words retrieved from titles and headings, abstracts, and keywords that appear 20 times or more).

Summary of the bibliometric analysis

AN IN-DEPTH ANALYSIS of the quantitative overview provided by the bibliometric review is somewhat difficult to condense. However:

- Science and Nature are the most influential journals, together with medicine, psychology, cognitive science, informatics and computer science.
- The combined area that we define as "sustainable AI" has, in the last 4-6 years, grown rapidly, but with an emphasis on the aforementioned aspect;
- American legal journals seem to be experiencing a broadening of knowledge and understanding. The method of analysis we have used, however, does not disclose whether the same legal scientific developments are occurring in Sweden or the Nordic countries.
- The most common concepts are "ethics", together with Big Data, AI and machine learning, unlike concepts such as "accountability" and "social bias", which occur less frequently.
- Data protection and privacy issues are relevant in a number of areas, and the co-citations analysis shows that AI and machine-learning are topics being discussed in the healthcare sector.

A narrower analysis of AI and machine learning in relation to ethics and delegation of responsibility is provided in Appendix 1.

III
In-depth studies

THE FOLLOWING SECTION consists of three in-depth studies intended to illustrate some of the practical implications, both with regards to potential benefits and challenges, that business sectors and other areas of focus face. We focus on: 1. Medicine: health and social care; 2. Telecom; and, 3. Digital platforms

Medicine: Future health care challenges

DEVELOPMENTS AND APPLICATIONS of new knowledge and technology in the healthcare sector are occurring at a rapid pace. Digitalisation, visualisation and simulations linked to AI, as well as applications and algorithms are causing a paradigm shift in the healthcare sector, and pose a challenge to medical ethics. New technologies in the healthcare sector that involve psychologically challenging situations, new kinds of interactions between humans and machines, and AI methodology is creating new challenges to the healthcare sector.

The human factor is often the direct cause of medical mistakes. But underlying safety issues and systematic errors should not be underestimated. These kinds of errors, which are seldom discussed, can be the result of a lack of standardisation as well as a lack of education and systematic training in critical thinking, as well as relevant knowledge and expertise. Access to vast data loads and information flows create new complexities in the healthcare system and challenges to human capacity.

Artificial intelligence and machine-learning in the healthcare sector

The hype surrounding AI and machine-learning has, by extension, changed our understanding of the seemingly endless potential of large amounts of information. The healthcare sector's earlier, cautious position has undergone rapid change, and the area is expanding quickly as a result of large amounts of data, processing power and innovation, which is related to a lag in much-needed, legal deliberations. In some, fundamental areas of medicine, such as medical image diagnostics, machine-learning has been proven to match or even surpass our ability to detect illnesses. An example of this is medical assessments of mammography images, and predicting lethal outcomes in the case of coronary artery disease. When designing algorithms, metrics and relevant references must be carefully defined if the algorithm is to work as planned. Systematic reviews of algorithms are also necessary, to be conducted in close collaboration with experts. The risks of bias and confounders must be managed since the original data on which the algorithms are based can lead to incorrect interpretations. Similarly, the models need to be optimised to avoid under-adaptation and over-adaptation.

Trust and accountability

Medical ethics and regulatory aspects also need to be managed. Delegation of responsibility in the event of failure needs to be clarified. There cannot be any ambiguity with regards to whether the designer of the algorithm or the individual using the algorithm for assisted decision-making is responsible. At the moment, this is a grey area in the health sector. Appropriate levels of understanding, transparency and oversight of self-learning, and decision-making applications must be

defined before they are commercialised, scaled up and implemented in the healthcare sector. The level of explainability and transparency required for the doctor to place trust in increasingly autonomous decision-making support systems must also be defined. Trust between caregiver and patient must be established as their relationship becomes increasingly dependent on third-party AI and machine-learning applications. Transparency is needed to allow oversight of commercial AI products and to ensure that they receive and can learn from appropriate feedback loops, and to delegate responsibility when products lead to undesirable and/or unexpected outcomes.

Conflicts of values and ethical challenges

There are other interests that also need to be weighed and considered. Machine-learning processes that are dependent on large amounts of data are often described as necessary components in the development of AI. This does, however, entail a conflict of values between rules that protect patient privacy and data, which, traditionally, has been a key aspect of medical ethics, and access to large amounts of patient data that AI applications in the healthcare sector require. How much attention should be paid to patient awareness and their wishes when their data is used to train AI applications? What about groups whose consent cannot be retrieved, e.g., because the databases used were conceived for a different purpose, or are out-dated, or are otherwise inappropriate? Another dilemma consists of problems that may arise when the patient does not understand the questions. For example, the Swedish Health and Medical Service Act sets out that healthcare efforts must be based on respect for the patient's right to self-determination and privacy. This often leads to a conflict between the benefits of innovation and data protection.

Discussions on future pricing of outcomes produced by algorithms compared to physical healthcare professionals is greatly needed. This is essential if we want future values, profits, costs and risks to be subject to transparency, and reasonable in relation to the required investments. These new challenges emphasise the need for a new approach to the healthcare sector in the future. Additionally, training programs in the

future must allow for evaluations and management of data and assisted decision-making applications before they can be implemented commercially and scaled up. Ethical questions and issues of responsibility must be raised to avoid them being left to drown in the wake of technological progress. Ethics have always been a key part of medical science and practical implementation in the healthcare sector. A new paradigm is required with regard to artificial intelligence and machine-learning.

Three examples that highlight challenges to applied AI in the healthcare sector

1. Large investments and marketing projects have recently been made in contraceptive apps. These fall under the category of medical devices, and the Swedish Medical Products Agency has regulatory authority in this area, but is not responsible for medical approval.

The app allows the (female) user to register her body temperature on a regular basis. Once the data has been submitted and calibrated, the app assesses the likelihood of pregnancy based on changes in body temperature related to fertility periods in the menstrual cycle. Following hundreds of complaints in connection with unwanted pregnancies, the Swedish Medical Products Agency has scrutinised and reviewed the app. The review was concluded and the app was approved after updates to the user instructions were made that clarified the risk of unexpected pregnancy. The Swedish Medical Products Agency also concluded that "private persons that are concerned, or have questions, about what contraceptives to use, should contact their healthcare representative for help and advice from a midwife or doctor".[68]

By updating the user instructions, the user's responsibility, and transparency with regards to the risk of unexpected pregnancy, was clarified. A clear ethical dilemma arises when the user is unable to understand or interpret the user instructions.

68 https://lakemedelsverket.se/Alla-nyheter/NYHETER---2018/Lakemedelsverkets-granskning-av-Natural-Cycles-avslutad/

2. Algorithm for estimating remaining lifetime of cancer patients with bone metastasis.[69] The goal is to reduce bias in treatment decisions regarding cancer patients with bone metastasis. This makes it easier to avoid under-treating or over-treating this category of patients. The product has been validated among different populations to increase its validity. Patient data is fed in and an estimated length of survival is produced. There is a possible ethical dilemma as well as a transparency issue that can arise when informing the patient and relatives. How does one explain and defend the choice of treatment to the patient and relatives if the chance of survival conflicts with the estimate?
3. Digital doctors for "triage". This application allows virtual, around-the-clock health check-ups via a chat bot. The bot asks questions and provides guidance with support from a decision tree. Sometimes, insurance companies begin by referring patients to this service as part of their health care package, since it is cheaper. While there is great potencial efficiency and scalability in this type of *platformisation* of health care, a general problem with this application is the risk of bias and confounders in the data on which patient recommendations are based. The risk of under-diagnoses, misdiagnoses, and over-diagnoses, all of which are potentially serious matters in terms of accountability, and with ethical concerns. Over-diagnoses can lead to unnecessary health care consultations "just in case", with escalating health care costs for potentially benign, self-healing conditions, at the expense of seriously ill patients.

Telecom

TELECOMMUNICATION TECHNOLOGY (TELECOM) enables AI to be broadly applied to many areas of society and business sectors. It allows for efficient gathering of data using sensors which is then aggregated to services and systems that request the information. This includes data that represents physical quantities in our environment (temperature, UV radiation,

69 https://www.pathfx.org/

noise levels, speeds, image and sound data, etc.) as well as information on how to use different kinds of items and services (digital content, repositioning, operating mode, online connections, etc.). In addition to transference of information between two units (voice conversations, video, etc.), these data points also provide us with the digital representation of the physical world as well as knowledge of how the physical and the digital worlds interface and are used over time. In more or less autonomous systems, feedback can then be looped back to the actuators within the networks that then implement decisions made with the assistance of AI systems, as well as provide feedback and connectivity between the various subsystems connected to the intelligent infrastructure.

With the large number of users, services and connected sensors globally, enormous amounts of data are managed in real-time around-the-clock. To make this possible, the telecom industry uses technology such as machine-learning and AI for automated decision-making.

Telecom companies currently operate networks in order to provide mobile communication for 5.7 billion subscribers around the world[70]. Companies that offer communication and online services require not only access to a communications infrastructure but also basic data on how individual users (or appliances) use the networks. This is partly in order to charge for the company's services, and also to allow them to optimise the networks' real-time capacity and accessibility. Networks that are constantly monitored by other systems to prevent serious operational disruption also emerge. These kinds of systems are growing increasingly autonomous as a result of AI. The goal is to be able to quickly identify network components that indicate errors, and, if possible, address the errors without the need for human intervention. Potential shifts of workloads in the network can be predicted, thereby allowing resource redistribution in order to maintain a high level of quality and ability when transferring data. Furthermore, information traffic within the systems are monitored in order to ensure that data that is transmitted through the network is

70 Ericsson Mobility Report https://www.ericsson.com/en/mobility-report/reports/november-2018

managed safely and with strong privacy protection. Trust in the system is also based on whether attacks and data theft can be prevented, and if possible, avoided in an efficient manner.

To manage telecom systems with assisted decision-making, certain challenges must be addressed. Some of the challenges described in more detail above, and some of the examples used to illustrate the challenges, directly relate to IT and the telecom sector.

However, there is yet another level of complexity that results from the fact that AI-based tools used in healthcare, transportation systems, the exercise of authority, etc., are often integrated with functions and characteristics of the intelligent infrastructure offered by the telecom sector to communicate data. This raises the question[71] of how to delegate responsibility when developing AI-based systems. Business organisations such as the IEEE[72] and standards organisations such as the ITU[73] have studied the question and initiated public discussions on the ethical aspects of these technological developments.

Digital platforms

DIGITAL PLATFORMS DO not represent a specific industry, but are, rather, an organisational model that is increasingly affecting a number of markets, and are being studied by more and more communications experts, legal scientists and economists.[74] This applies, not least, to challenges arising from content moderation, which to a large degree deals with

71 https://www.linkedin.com/pulse/nordic-engineers-send-message-politicians-inese-podgaiska/

72 https://ethicsinaction.ieee.org

73 https://medium.com/@UNDP/who-is-writing-the-future-designing-infrastructure-for-ethical-ai-4999620db295, https://news.itu.int/challenges-and-opportunities-of-artificial-intelligence-for-good/

74 van Dijck, J.; T. Poell & M. de Waal (2018). The Platform Society. Public Values in a Connective World. Oxford University Press; Andersson Schwarz, J. & Larsson, S. (red. 2019) Plattformssamhället. Den digitala utvecklingens politik, innovation och reglering. Stockholm: Fores; Larsson, S. & Andersson Schwarz, J. (eds. 2018) Developing Platform Economies. A European Policy Landscape. Brussels and Stockholm: European Liberal Forum asbl and Fores.

automating detection and policy implementation by way of applications such as image recognition and language analysis.[75] One particular challenge lies in the combination of normative, i.e., value-based, choices together with automated processes that learn from human expressions, values and social structures. In automated decision-making processes, transparency, as mentioned previously, is a multi-faceted issue[76]. Questions of accountability for how algorithms are designed and implemented in platforms needs much further scrutiny in how they interact with social structures that risk being inherently biased in themselves, i.e., discriminatory, racist, hateful, and where different groups may disagree wildly on normative preferences.

It can be concluded that moderating large-scale platforms that involve cultural and medial expressions becomes extremely complicated with regards to where to draw lines, in contextual conflicts, with political opponents that "flag" their interlocutors to obstruct them, or other abuses of the design of the platforms, in addition to unintended effects that result from automated policies. Problems concerning not only what should be deemed as unacceptable behaviour, but also different cultural attitudes, conflicting legal orders and sensitive issues emerge quickly, as media researcher Gillespie states:

> ... balancing offense and importance; reconciling competing value systems; mediating when people harm one another, intentionally or otherwise; honoring contours of political discourse and cultural taste; grappling with inequities of gender, sexuality, race, and class; extending ethical obligations across national, cultural, and linguistic boundaries; and doing all that around the hottest hot-button issues of the day.[77]

75 Gillespie, T. (2018a). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

76 Larsson, S. (2018) "Sju nyanser av transparens: Om artificiell intelligens och ansvaret för digitala plattformars samhällspåverkan," i Andersson Schwarz, J. & Larsson, S. (red.) Plattformssamhället. Den digitala utvecklingens politik, innovation och reglering. Stockholm: Fores.

77 Gillespie (2018, p. 10).

Transparency in AI processes is a core issue for digital platforms. Transparency must be constantly weighed against proprietary arrangements, a centralist approach to how the platform is steered, and scalability involving large numbers of simultaneous users. The size of these large-scale digital platforms and the extent of data-driven automation makes their impact on society, and their inherent risks, highly tangible. Unintended consequences of the platforms' normative efforts to implement automated policies are a key issue when it comes to understanding infrastructural and market-creating designs that have an impact on entire markets, individuals and companies.

One normative challenge lies in when machine-learning and other kinds of AI comprise core processes in applications that interact with people and social structures, since they also interact with structural and social biases. There is often a lack of a neutral, normative understanding; should an application passively reproduce social biases or should it actively and normatively counteract them? Should imbalances in gender, power, ethnicity, economy, and religion be challenged, or could they be used to determine relevance assessments, pricing, etc.? And, if we decide that they should be challenged, the norms by which they are judged must also be defined – whose norms should take precedence for platforms spanning both multiple jurisdictions and cultures? How, and by whom, do we go about defining them?

AI and machine-learning have the potential to be of use for platforms in relation to this aspect of digitalisation. At the same time, as digital platforms become increasingly important in our daily lives, these data-driven technologies raise new questions of fundamental importance to society with regards to accountability, appropriate levels of transparency, and, importantly, how to create trustworthy autonomous and artificial decision-making processes.

Appendix 1:

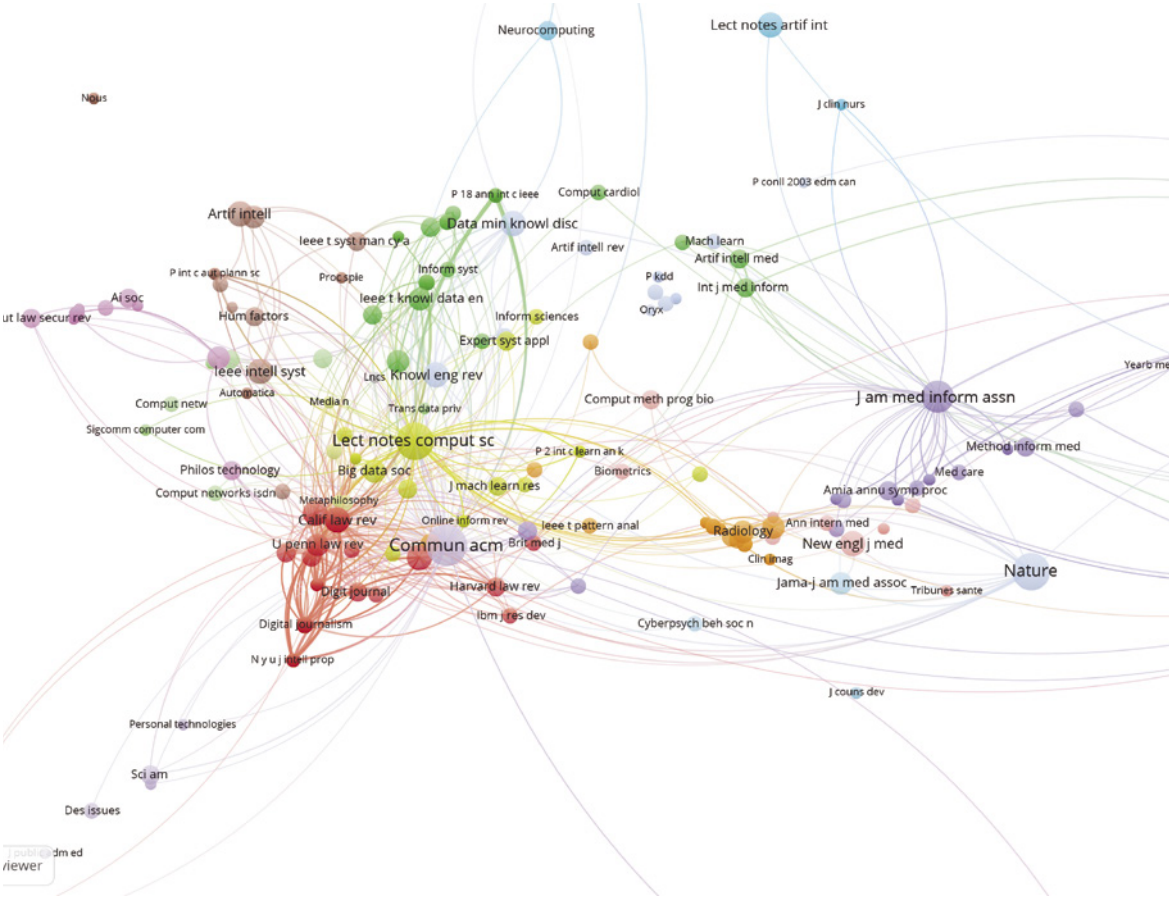
Peer-reviewed articles in general

A sample excerpted from the Web of Science databases, with a particular focus on ethics and delegation of responsibility, in combination with either AI or ML, resulted in 46 articles after co-citation analysis, as follows:

"artificial intelligence" OR "machine learning"

AND

"ethic" OR "accountability"


Neurocomputing
Lect notes artif int
Nous
J clin nurs
P conll 2003 edm can
P 18 ann int c ieee
Comput cardiol
Artif intell
Data min knowl disc
Ieee t syst man cy a
Artif intell rev
Mach learn
Artif intell med
P int c artif plann sc
Proc spie
Inform syst
P kdd
Int j med inform
Ai soc
Ieee t knowl data en
Oryx
Ut law secur rev
Hum factors
Inform sciences
Expert syst appl
Ieee intell syst
Lncs
Knowl eng rev
Automatica
Media n
Trans data priv
Comput meth prog bio
J am med inform assn
Comput netw
Sigcomm computer com
Lect notes comput sc
P 2 int c learn an k
Method inform med
Philos technology
Big data soc
Biometrics
Med care
Comput networks isdn
Metaphilosophy
J mach learn res
Amia annu symp proc
Calif law rev
Online inform rev
Ieee t pattern anal
Radiology
Ann intern med
U penn law rev
Commun acm
Brit med j
New engl j med
Clin imag
Nature
Digit journal
Harvard law rev
Jama-j am med assoc
Tribunes sante
Ibm j res dev
Digital journalism
Cyberpsych beh soc n
N y u j intell prop
J couns dev
Personal technologies
Sci am
Des issues

Sustainable AI

THIS REPORT IS an inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence conducted within the Swedish Vinnova-funded project "Hållbar AI – AI Ethics and Sustainability", led by Anna Felländer. Based on a review and mapping of reports and studies, a quantitative and bibliometric analysis, and in-depth analyses of the healthcare sector, the telecom sector, and digital platforms, the report proposes three recommendations. Sustainable AI requires: **1.** a broad focus on AI governance and regulation issues, **2.** promoting multi-disciplinary collaboration, and **3.** building trust in AI applications and applied machine-learning, which is a matter of key importance and requires further study of the relationship between transparency and accountability.

WWW.AISUSTAINABILITY.ORG