THE SOCIO-LEGAL RELEVANCE OF ARTIFICIAL INTELLIGENCE

Stefan Larsson
Lund University

AI SUSTAINABILITY CENTER

The Socio-Legal Relevance of Artificial Intelligence

Stefan Larsson, Lund University
2019

This is a preprint of an article in forthcoming special issue
"Le droit à l'épreuve des algorithmes", Droit et société, 103(3)
(Ed. by Dubois C. & Schoenaers F.)

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 organizations, 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 center operates with a research-based practical framework to help organizations sustain their values in their AI applications and how they are scaled in a broader ethical and societal context. The center aims to help both private and public organizations to both gain trust by transparency and to be ahead of the regulatory curve. These "best practice" from the center can support in the process of identifying the level of transparency and explainability in relation to the stakes it posed.

“Models are opinions embedded in mathematics.”

CATHY O’NEIL,
COMPUTER SCIENTIST AND AUTHOR OF THE BOOK,
WEAPONS OF MATH DESTRUCTION (2016)

Contents

STEFAN LARSSON IS a lawyer (LLM) and Associate Professor in Technology and Social Change at Lund University, Department of Technology and Society. He holds a PhD in Sociology of Law as well as a PhD in Spatial Planning. In addition, Dr. Larsson is a senior researcher and head of the Digital Society program at the Swedish think tank Fores and scientific advisor for the Swedish Consumer Agency as well as the AI Sustainability Center. His research focuses on issues of trust and transparency on digital, data-driven markets, and the socio-legal impact of autonomous and AI-driven technologies.

Among his publications:

Algorithmic Governance and the Need for Consumer Empowerment in Data-driven Markets
INTERNET POLICY REVIEW 7(2), 2018

Conceptions in the Code.
How Metaphors Explain Legal Challenges in Digital Times
OXFORD UNIVERSITY PRESS, 2017

Preface

IN TODAY'S DATA-DRIVEN era, AI has opened up numerous advancements and opportunities for people and society. At the same time, consumers and citizens increasingly trade data about their personal preferences and behavior in exchange for convenience and time efficiency. Many do not understand the intricacies of this "trade," and as a result we are seeing more controversial incidents from AI applications resulting for example in discrimination and privacy intrusion. The implication is that the need and urgency for transparency in AI applications and how they are governed has escalated. In this report, Larsson highlights important trade-offs and conflicts of interests, for example between transparency and privacy within healthcare, and the need for better understanding by regulators before introducing new regulation, and/or adopting existing regulations to the new environment. In order to address issues relating to cultural values, norms and ethics, Larsson argues for an interdisciplinary and multidisciplinary approach when striving for trusted and trustworthy AI-systems applied in society.

In his analysis, Larsson identifies what he calls the normative mirroring effect of using human values and societal structures as training data for learning technologies. That AI, learning from real world examples derived from human activities, can act as a mirror for social structures, reproducing not only the beneficial and desired but also the biased, skewed and discriminatory. This leads to a number of questions, including how to address accountability for those devising the mirror, signaling that the design of AI for some cases can be seen as normative rather than strictly

neutral. If so, should AI-systems strive to reproduce the world as it is or in accordance with a “preferred” reality? Larsson argues that the normative component of systems interacting with human values stresses the importance of multidisciplinary competence in development and deployment of such systems.

In order to embrace the value from AI and minimize the downsides, it is crucial for organizations to adopt appropriate levels of transparency, and for the regulators to understand that transparency has different nuances and competing interests, which is why Larsson’s research is extremely timely and important.

ANNA FELLÄNDER, AI SUSTAINABILITY CENTER
ELAINE WEIDMAN GRUNEWALD, AI SUSTAINABILITY CENTER
LI FELLÄNDER-TSAI, KAROLINSKA INSTITUTE
FREDRIK HEINTZ, LINKÖPING UNIVERSITY

Introduction: AI and Society

IN RECENT YEARS, the field of artificial intelligence (AI), in particular machine learning, has undergone significant developments. The underlying technologies and methods are useful in a number of applied areas and interactive spaces on markets and in society, and particularly useful in information-intensive and digitalized environments. For example, it can be used for automated differentiated pricing methods for hotel bookings and airline tickets, for targeted and personalized marketing online and in loyalty card systems, for individual relevancy in search engines, music recommendation systems or understanding and replying in voice conversations. Our homes are increasingly becoming equipped with self-learning thermostats, other "property technology" and virtual assistants embodied in smart speakers. AI is also being applied directly to actual life or death matters. Currently, self-driving cars and other vehicles with various degrees of autonomy are under development, as are AI-assisted tools used for cancer diagnoses, predictive risk-analyses produced by insurance companies and creditors, image recognition algorithms used in social media, police enforcement and security services, or for military purposes, such as drones developed for remote warfare.

Drawing from socio-legal concerns of what digital and increasingly autonomous technologies means for law and society,[1] this article outlines

1 Stefan LARSSON. "Sociology of Law in a Digital Society – A Tweet from Global Bukowina", *Societas/Communitas* 15(1). 2013, p. 281-295; cf. Bourcier, Danièle. "De l'intelligence artificielle à la personne virtuelle: émergence d'une entité juridique?" *Droit et société* 3, 2001, p. 847-871.

some of the legal and societal challenges that the use of AI and machine learning entails. Specifically, the main argument is focusing normativity in design, societal bias in autonomous and algorithmic systems, as well as difficulties with distribution of liability and accountability. In addressing the close relationship between accountability and transparency, the article proposes seven "nuances" or aspects of transparency, suggested as a socio-legal contribution to the already present notion of *explainability* within AI research *(XAI)*.[2] Thus, the focus in this article is not primarily on clearly defining what AI is according to a computer scientific perspective, but on pointing out the social significance of an everyday and practically applied AI from a socio-legal perspective, stressing the need for keeping society "in-the-loop".[3] This is of key importance from the perspective of defining what technological advancements and applications are to be seen as fair and normatively just – which arguably should be seen as a continuous assessment. In addition, and perhaps of particular socio-legal value, this is of key importance also from the perspective that self-learning and autonomous technologies that depend on data that is derived from human values, behaviours and social structures will not only face and reproduce the balanced sides of humanity, but also the biased, skewed and discriminatory. This represents a sort of mirroring effect with great normative implications for designers and developers, that I develop further below.

In conjunction with society's increasing use of, and dependence on, AI and machine learning, there is indeed a growing societal need to understand potentially negative consequences and risks, how various interests and power are distributed, and what kinds of legal and ethical frameworks, standards, certifications or procedural stances might become necessary. Literature that deals with artificial intelligence endowed with different levels of autonomy and agency has a long tradition of formulating rules and normative principles. Perhaps the most famous ones

2 Or Biran & Courtenay Cotton. "Explanation and justification in machine learning: A survey". In *IJCAI-17 Workshop on Explainable AI (XAI)*, 2017.

3 Cf. Iyad Rahwan, "Society-in-the-loop: programming the algorithmic social contract", *Ethics and Information Technology* 20(1), 2018, p. 5-14.

are Isaac Asimov's three laws of robotics from 1942, later followed by a number of others within the field of robotics research.[4] In earlier years, any concerns about regulation and ethics often pertained to an imagined, somewhat unspecified form of artificial intelligence that could, based in its instinctual and analytical capacity, *revolt against* humanity. Today, such concerns are sometimes expressed in terms of a potential, future super-intelligence, and a fear that technological progress could lead to an upgradable and self-improving artificial intelligence – a sort of "singularity" in which humanity, as we know it, basically becomes extinct.[5]

This article does not, however, focus on a perceived super-intelligence or general artificial intelligence, but rather, on contemporary, everyday versions of artificial intelligence in order to relate them to relevant legal and socio-legal challenges. Therefore, in this article I adopt a broad definition of AI that covers a number of technologies and analysis methods, such as machine learning, natural language processing, image recognition, neural networks and deep learning. In particular, machine learning, which, briefly put, deals with how to "teach" computers to learn from data without having to specifically programme computers for that particular task, is a field that has developed at an extremely rapid pace in recent years as a result of a vast, historically incomparable accumulation of data and greatly increased analytical processing power. Although the term "machine learning" was coined in 1959,[6] the field has progressed from being a sub-discipline with the ambition to develop artificial intelligence to being applied to solve practical problems, with a focus on predictive analyses based in training data. Today, this area is generally included in the field of artificial intelligence, but it is also closely linked to statistics and image recognition, where machine learning has proven to be highly useful in a number of practical applications. A key component of AI in general, and machine learning in particular, is the algorithms used, developed and studied to create software with the capacity to learn and

4 Susan Leigh Anderson. "Asimov's 'three laws of robotics' and machine metaethics", *AI & Society* 22(4). 2008, p. 477-493.

5 Cf. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

6 Arthur Samuel. "Some Studies in Machine Learning Using the Game of Checkers", *IBM Journal of Research and Development* 3(3). 1959, p. 210-229.

produce probability assessments. The main difference between earlier AI-related rules and ethical principles and contemporary times is that, today, discussions on how they should be regulated now concern everyday uses of AI and machine learning in a digitalized and increasingly data-driven reality. The starting point, here, is that a number of social practices – which have an impact on working life, ordinary families' financial situation, the dissemination of news and knowledge and healthcare issues – are now mediated using artificial intelligence. This raises a number of questions that need to be examined from a socio-legal perspective and which are studied trisectionally in this article:

1. How can fairness in AI be understood from a socio-legal perspective? E.g. which social norms are reproduced or strengthened by self-learning, autonomous systems, and how does normativity relate to data-dependent AI?
2. How can issues of accountability with regards to applied AI be problematized from a socio-legal perspective, e.g. in relation to increasingly autonomous applications, artificial agents and automated decision-making?
3. What are the key interests at play in transparent and explainable AI, from a multidisciplinary and socio-legally informed perspective? This relates to a balancing of not necessarily compatible interests, how society could or should supervise AI applications and their implications, and how to formulate explanations, insights and knowledge with regards to these applications.

The purpose here is to contribute to a broad, legal and socio-legal orientation by describing some of the legal and normative challenges posed by applied AI. Recently, political discussions in many countries as well as the EU have begun to address the challenges facing regulatory efforts in data-driven markets, and in particular, algorithm-driven developments in machine learning and artificial intelligence. In December 2018, the EU Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG)

published a draft of ethics guidelines for trustworthy AI,[7] that resulted in a final publication after consultation, in April 2019.[8] In May 2018, the Swedish government, for example, published the National Approach for Artificial Intelligence (*Nationell inriktning för artificiell intelligens),* which, among other things, includes a section on the need for Sweden to "develop rules, standards, norms and ethical principles to guide ethical and sustainable AI, and the use of AI".[9] From a theoretical standpoint, this terminology raises several questions regarding how to distinguish between and define these concepts and their practical implications; however, they should be interpreted as expressing a need to impose some form of restrictions on the development and implementation of a powerful, potentially independent, opaque and complex technology in core social functions and markets.

7 European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) "Draft Ethics guidelines for trustworthy AI," 18 December 2018. https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai

8 AI HLEG. *Ethics Guidelines for Trustworthy AI.* Brussels: The European Commission. 2019.

9 Regeringskansliet. *Nationell inriktning för artificiell intelligens*. Näringsdepartementet, 2018, p. 10.

Socio-Legal Challenges of AI: FAT

WHEN IT COMES to data, algorithm-driven systems, and the potential social consequences of artificial intelligence, a growing understanding of the importance of legitimacy, fairness, ethical and human-centric approaches, is emerging in the literature. A relatively new field, therefore, has come to focus on *Fairness, Accountability* and *Transparency,* abbreviated as FAT.[10] Research in field emphasizes that algorithmic systems are used in many situations where vast amounts of "Big Data" are implemented to filter, categorize, rate, recommend, personalize, and in other ways shape human experiences and relations. Although these systems have many benefits, they also carry inherent risks, such as the codification and reinforcement of social prejudices, diminished responsibility and increased asymmetry of information between the data producers (i.e., the customers) and data owners.

At the same time, this relatively new concept (FAT) addresses issues that have long been the subject of research in the social sciences and the humanities, i.e. ethical and philosophical theorizing. Transparency, with its conceptual history, is often seen as a fundamental cornerstone of supervision and vital component of achieving accountability.[11] Also, issues of "fairness" may

10 E.g., see https://www.fatml.org; For an overview of research on ethical, social and legal consequences of AI, see Stefan LARSSON, Mikael ANNEROTH, Anna FELLÄNDER, Li FELLÄNDER-TSAI, Fredrik HEINTZ, and Rebecka CEDERING ÅNGSTRÖM. *Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence*. Stockholm: AI Sustainability Center, 2019.

11 For an analysis on the conceptual origins and background of 'transparency' with regards to AI, see Stefan LARSSON & Fredrik HEINTZ, "AI Transparency", *Internet Policy Review*, forthcoming.

draw from a rich literature on justice and normativity, knowledge based in the broader, empirically based legal science of sociology of law.

Fairness

There are a number of examples where unintended social prejudices are reproduced or automatically strengthened by AI systems which often only become apparent following rigorous study. A few examples:

- Computer science researchers at the University of Virginia discovered that some popular image databases had a gender-based bias which portrayed women in the kitchen and men out hunting, resulting in a machine learning application that not only reproduces but also reinforces these biases.[12]
- A critical article by investigative journalists at ProPublica[13] that focuses on the American authorities' use of algorithm-guided practices based on recidivism predictions, i.e., the probability of relapses into crime, showed that the so-called COMPAS system was more likely to *incorrectly* predict increased crime rates among black offenders while simultaneously, and incorrectly, predicting the opposite where white offenders were concerned.[14]

12 As reported in Wired, "Machines taught by photos learn a sexist view of women", by Tom Simonite, 21 August 2017: https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/amp ; for a study, see Jieyo Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, & Kai-Wei Chang. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints". *arXiv preprint*, 2017, arXiv:1707.09457.

13 The study was carried out and published by civil rights-motivated investigative journalists at ProPublica, "Machine Bias", by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. 23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

14 This case is discussed in a growing body of literature from several angles, and is particularly interesting from a socio-legal perspective, not the least from the fact that it is explicitly dealing with the automation of court decisions; cf. Robyn Caplan, Joan Donovan, Lauren Hanson, and Jeanna Matthews (2018). *Algorithmic Accountability: A Primer*, NYC: Data & Society. 2018. For a critique of the judicial use of automated risk assessment tools in ways that undermine the fundamental values of due process, equal protection and transparency, see Han-Wei Liu, Ching-Fu Lin, and Yu-Jie Chen. "Beyond State v Loomis: artificial intelligence, government algorithmization and accountability." *International Journal of Law and Information Technology* 27(2): 122-141, 2019.

- In an effort to improve transparency in automated marketing distribution, a research group developed a software tool to study digital traceability and found that such marketing practices had a gender bias that mediated well-paid job offers more often to men than to women.[15]
- A study of three commercial, gender-based image recognition systems showed that the most incorrectly categorized group consisted of dark-skinned women.[16] This means, among other things, that their services, and the applications based on them, work poorly for people with certain physical characteristics. Also, there is a significantly narrower margin of error when it comes to white males.

The term "bias" is also used in statistics and computer science and therefore has several different meanings, which means that there is some confusion surrounding this term which might complicate social scientific and techno-scientific understandings of the concept.[17] In the present context, I will use the term "social bias", based in a socio-legal understanding of social norms and cultural values.

Value–based discussions surrounding machine learning and AI are often conducted in terms of "ethics", as in the report *Ethically Aligned Design,* published by the global technical organization IEEE.[18] Such discussions on the topic of "ethics" and artificial intelligence, in this context, reflect a broad understanding that we as a society need to reflect on values and norms in AI developments, as well as – and this understanding is gaining force in social scientific literature – the impact AI is having on us, on society, and the values, culture, power and opportunities that are

15 Amit Datta, Michael Carl Tschantz, Anupam Datta. "Automated Experiments on Ad Privacy Settings – A Tale of Opacity, Choice, and Discrimination". *Proceedings on Privacy Enhancing Technologies*. 1, 2015, p. 92–112, DOI: 10.1515/popets-2015-0007.

16 Joy Buolamwini, & Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency,* 2018, pp. 77-91.

17 As noted by, among others, Arvind Narayanan, A. "21 fairness definitions and their politics", presented at the conference on *Fairness, Accountability, and Transparency*, 2018. http://fairmlbook.org/tutorial2.html

18 The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. IEEE, 2019.

reproduced and reinforced by autonomous systems. Therefore the use of the concept of "ethics" in contemporary AI governance discourse may arguably be seen as a kind of proxy; i.e., it represents a conceptual platform with the capacity to bring together the diverse groups that develop these methods and technologies – i.e., mathematicians and computer scientists – with groups that commercialise and implement them in the market, as well as those groups that study these methods and technologies and their role in in society from a social scientific and humanities-oriented perspective, in order to gain a better understanding of their impact. Discussions on ethics in AI will, in time, likely be replaced by more clearly defined concepts in the areas of regulation, industry standards, certifications, and more in-depth analyses of culture, power, market theory, norms, etc., in the main areas of traditional scientific fields. For many years, sociologists of law have studied legitimacy in terms of social norms, in line with Durkheims "social facts"[19] or Erlich's "living law",[20] Pound's "law in action",[21] which see social norms as an object that can be empirically measured, is structurally widely dispersed, but has not necessarily been formalised in terms of law "in books".[22]

The fact that computerised systems may be biased or have socially problematic or one-sided cultural values is not necessarily new knowledge,[23] but the rapid development of such systems in conjunction with society's dependence on them is, now, greater than ever, and has consequences for key social functions, such as credit rating, employment opportunities,

19 Émile Durkheim (1982) [1st pub. 1895]. Steven Lukes (ed.). *The Rules of Sociological Method and Selected Texts on Sociology and its Method*. W. D. Halls (translator). New York: Free Press; Cf. Roger Cotterrell. *Emile Durkheim: Law in a Moral Domain*. Edinburgh University Press, 1999.

20 Eugen Ehrlich. *Fundamental Principles of the Sociology of Law*. New Brunswick, NJ: Transaction Publishers, 2002. For a modern application, see for example Rustamjon Urinboyev and Måns Svensson. "Living law, legal pluralism, and corruption in post-Soviet Uzbekistan." *The Journal of Legal Pluralism and Unofficial Law* 45(3), 2013, p. 372-390.

21 Roscoe Pound. "Law in books and law in action." *American Law Review*, 44:12, 1910.

22 E.g. Håkan Hydén & Måns Svensson, "The concept of norms in sociology of law." In: Wahlgren P. (ed.) *Scandinavian Studies in Law*. Stockholm: Law and Society, 2008, pp. 15–33; Måns Svensson & Stefan Larsson, "Intellectual Property Law Compliance in Europe: Illegal File sharing and the Role of Social Norms", *New Media & Society*, 14(7), 2012, p. 1147-1163.

23 Cf. Batya Friedman & Helen Nissenbaum. "Bias in Computer Systems," *ACM Transactions on Information Systems*, 14(3). 1996, p. 330-347.

health care issues, and the dissemination of knowledge and news.[24] For example, an analysis on two large, publicly available image data sets found that these exhibit what was called an observable "amerocentric and eurocentric representation bias".[25] That is, they were skewed towards cultural expressions in the western world, resulting in lack of precision for expressions in the developing world. Furthermore, social, political, economic and cultural aspects of search engines, for example, have been the subject of a large number of studies,[26] as have the cultural implications of policies on obscene or taboo language and so-called "auto-complete" functions used by search engines, i.e., the function that allows search engines to fill in additional information, which can sometimes lead to controversial results.[27]

Recently, American Professor of Information Science Safiya Noble (2018) strongly underlined, in her book, *Algorithms of Oppression: How search engines reinforce racism,* that search engines, which are largely automated and have self-learning and artificial intelligence characteristics, interact, reproduce and are a product of social, historical and cultural structures. Therefore, algorithms can automatically limit the opportunities available to individuals in a way that may be unlawful, or could be considered unethical. This implies a sort of "technological redlining", to use Noble's term, in which data-analyses opaquely and structurally discriminate against certain groups, and which is often only observable through extensive study after the event. The terminology is inspired by the "redlining" popularized in the US in the '60s to describe a discriminatory practice of highlighting areas (in red on a map) that banks should avoid investing in based on social demographics, and the term has also been used to describe systematically weakened access to financial

24 Cf. Larsson et al., 2019; Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, Oscar Schwartz, *AI Now Report 2018*. New York, 2018.

25 Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. "No classification without representation: Assessing geodiversity issues in open data sets for the developing world." *arXiv preprint arXiv:1711.08536 (2017).*

26 Cf. Eszter Hargittai. "The social, political, economic, and cultural dimensions of search engines: An introduction". *Journal of Computer-Mediated Communication*, 12(3), 2007, p. 769-777.

27 Rex L. Troumbley. *Taboo language and the politics of American cultural governance*. Doctoral dissertation, University of Hawai'i at Manoa, 2015.

services, insurance, health care services, etc., in certain neighbourhoods.[28] Noble uses the term to underline the responsibilities of digital intermediaries that interact with – and thereby contribute to – already existing discrimination practices.

Thereby, Noble connects technological redlining to a long history of prejudice that is now being transferred to a technological *datafied* context. This lack of overview and transparency poses a challenge, because these methods are "increasingly elusive because of their digital deployments through online, internet-based software and platforms, including exclusion from, and control over, individual participation and representation in digital systems".[29] Therefore, there are consequences to technological redlining when individuals subject to such profiling have no control over how their personal data is used. If the data contains social bias, it becomes reproduced in the profiling results. In the absence of applicable mechanisms to ensure transparency or review how the data is used or delegate an appropriate level of responsibility, it becomes extremely difficult, Caplan et al. argue, to gain an awareness of algorithmic decisions that lead to obstacles or limits on civic rights.[30] This means that there is a need of greater transparency in the application of data-driven autonomous services and platforms.

Systems that reproduce bias have also been criticized from the standpoint that an overly homogeneous design community leads to blind spots. For example, a report by AI research centre AI Now on "legacies of bias" argues 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

28 It is sometimes attributed to American sociologist John McKnight, cf. William Norton. *Cultural Geography: Environments, Landscapes, Identities, Inequalities*. Oxford University Press. 2013. A number of studies suggest a long-standing relationship between geography, race and contemporary housing and credit markets; cf Jesus Hernandez. "Redlining revisited: mortgage lending patterns in Sacramento 1930–2004." *International Journal of Urban and Regional Research* 33, no. 2. 2009, p. 291-313.

29 Noble in Caplan et al., 2018, p. 4.

30 Caplan et al. 2018.

> for change. Machine predictions and performance are constrained by human decisions and values, and those who design, develop, and maintain AI systems will shape such systems within their own understanding of the world. Many of the biases embedded in AI systems are products of a complex history with respect to diversity and equality.[31]

In line with this, one may conclude that values and normativity can be found on both sides of the design process; i.e., in the use of structurally biased data retrieved from individuals and society, as well as in the design and development of applications and services. This prompts complex but necessary questions of who is to be held accountable for what in autonomous systems applied in society.

Agency and Accountability

There are several, parallel approaches to questions of accountability in the context of AI. *Agency,* it seems, is one of the crucial parts. An important aspect of the delegation of legal responsibility deals with assessments of intentions, expectations and knowledge of the risks of certain activities.[32] Can a machine or software "understand" things and have "intentions"? These questions might not be relegated to a distant future, and regardless of the answers, these discussions will have legal implications, as companies and authorities develop increasingly autonomous AI services that will unavoidably be subjected to judicial proceedings. These might range from discriminatory outcomes of large scale automated decision-making to car accidents involving self-driving cars, or unexpected costs related to smart thermostats.

A governance approach on AI expressing principles or guidelines has a long tradition but comes with a newfound vigour. Conventional AI

31 Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker & Kate Crawford. *AI Now 2017 Report*. AI Now Institute at New York University. 2017, p. 18.

32 Mireille Hildebrandt. *Smart Technologies and the Ends of Law*, UK & USA: Edward Elgar Publishing, 2015.

research has, as mentioned, previously referenced Asimov's robotic laws,[33] and business organizations and research groups have developed a series of principles for robotics and machine learning. Some companies have also laid out principles for their AI development projects. The aforementioned IEEE report focuses on responsibility issues from a design and designer perspective, and also discusses autonomous weapons as a particularly problematic field. In June 2018, Google set out a handful of principles for artificial intelligence[34], just a few weeks after it had become known that the company had decided not to renew their Maven project[35] contract with the American armed forces, which focused on developing machine learning to analyse drone videos. A large number of researchers in the field have begun to express a growing awareness of harmful and malicious implementations of AI that also addresses the responsibilities of those involved in design and development.[36] The threat, here, has to do with, among other things, the development of different methods of cyber-attacks, such as automated hacking and online, remotely controlled, autonomous vehicles which could be used in physical attacks, e.g., by steering them into crowds. This also includes the use of politicised and polarising bot networks to influence elections, as in the run-up to the Brexit election,[37] or to disrupt various social issues, such as public discussions on vaccinations in the USA.[38] From a security perspective, the field of research that studies malicious uses of AI has called for AI development teams to adopt a culture that takes more responsibility for their tools and how they can

33 Leigh Anderson. 2008, p. 477-493.

34 Sundar Pichai. "AI at Google: our principles", Google blog, 7 June, 2018. https://www.blog.google/topics/ai/ai-principles/

35 The Verge. "Google reportedly leaving Project Maven military AI program after 2019", by Nick Statt. Jun 1, 2018. https://www.theverge.com/2018/6/1/17418406/google-maven-drone-imagery-ai-contract-expire [last visited 10 June 2019].

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

37 Marco, T. Bastos & Dan Mercea. "The brexit botnet and user-generated hyperpartisan news." *Social Science Computer Review*, 2017. https://doi.org/10.1177/0894439317734157

38 E.g., David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. "Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate", *American Journal of Public Health*, 2018. DOI: 10.2105/AJPH.2018.304567; For more on the social impact of platforms, see Stefan Larsson & Jonas Andersson Schwarz, *Developing Platform Economies. A European Policy Landscape*. Brussels and Stockholm: European Liberal Forum asbl and Fores, 2018.

be used, and emphasizes the importance of education, ethical standards and norms.[39]

It is often argued, in critical discussions on the impact of algorithms, that the risk of bias being recurrently automated and injected into processes is a key challenge – even when the intent is not conscious, malicious abuse. As mentioned, this can occur as a result of training data that is one-sided, outdated or otherwise poorly represents the desired outcome.[40] Caplan et al. refers "algorithmic accountability" to the process of delegating responsibility for damages resulting from algorithmically controlled decision-making that leads to discriminatory or unfair consequences.[41] Such accountability could also address responsibility issues with regards to how algorithms are developed, and their impact on, and consequences for, society. In the event of any harmful effects, responsibly managed systems should be equipped with mechanisms that allow for reparative measures.

> While law has always lagged behind technology, in this instance technology has become de facto law affecting the lives of millions – a context that demands lawmakers create policies for algorithmic accountability to ensure these powerful tools serve the public good.[42]

This statement echoes legal scholar Lawrence Lessig's arguments over a decade ago that "code is law" and that the actual digital architecture itself must be included when analysing norms and behaviours.[43] However, AI, it seems, comes with an additional layer as the code does not single-handedly reveal what steering model is being developed when a machine learning algorithm is analyzing patterns in large sets of data. Code – and its analytical and "learning" data processing – may lead to the informal coded laws Lessig formulated, the digital architecture governing

39 Brundage, M. et al. 2018, p. 7)

40 Cf. Engin Bozdag. "Bias in Algorithmic Filtering and Personalization". *Ethics and Information Technology* 15 (3). 2013, p. 209-227.

41 Cf. Nicholas Diakopoulos. "Algorithmic accountability: Journalistic investigation of computational power structures". *Digital Journalism*, 3(3). 2015, p. 398-415.

42 Caplan et al., 2018, p. 12.

43 Lawrence Lessig, "Code is law." *The Industry Standard* 18, 1999; Lawrence Lessig. *Code: Version 2.0.* 2006; Cf. Stefan Larsson, 2013.

automated decisions, today on digital platforms influencing billions. This is a newfound AI-driven architecture layered on top of the code Lessig likely was aiming for originally, but his core argument remains intact, that we need to understand how the code regulates, what values that emerge from it. A major shift, however, from the 15-20 years that has passed since the inception of those ideas is that the Internet has gone through fundamental changes, from a highly distributed non-professional web to one highly moderated by a fewer set of gigantic digital platforms.[44]

Another related, inherent challenge has to do with making future predictions: i.e., machine learning applications that can be used to make probability assessments of events that have not yet occurred. How serious a problem this poses – what stakes that are involved – depends on what such assessments are used for. If a probability assessment is used, for example, for credit rating purposes, medical diagnoses, delegation of law enforcement resources or penal recommendations, it is surely underlining the extreme importance of ensuring that the prediction is as fair and auditable as possible.

To demonstrate how AI and machine learning have become components of complex areas in society which further highlight the need to recognize AI as a social challenge, two examples can be mentioned, here: digital platforms and autonomous vehicles.

Digital Platforms

Further elaboration on the problems of delegating responsibility in an AI context leads us to study the important role of digital platforms, which unavoidably brings up the issue of how to assess the responsibilities of intermediary actors for contents or behaviours that are disseminated or generated via platforms. Questions concerning the responsibility

44 Cf. Jonas Andersson Schwarz, "Platform Logic: An Interdisciplinary Approach to the Platform-Based Economy". *Policy & Internet* 9(4). 2017, p. 374-394; Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

of intermediaries are nothing new,[45] but contemporary examples can be found in large-scale digital platforms, e.g., in discussions on the responsibilities of Facebook and YouTube (i.e., Google) for information shared between their platforms, and whether Google's search engine indexing makes relevance assessments.[46] Since these are large-scale platforms – Facebook has over 2 billion active users and Google is reported to provide no less than seven services that are used by over 1 billion users – they automate their information management processes to a high degree. Both operators are major investors in, and developers of, artificial intelligence for a number of functions, such as facial recognition, language analysis and voice recognition, etc.[47] One variation of the question concerning the responsibility of intermediaries deals with the level of control of user information, as highlighted in the so-called Cambridge Analytica scandal, where between 50 and 87 million Facebook users' personal details where used to influence democratic elections in a number of countries.[48] When Facebook's CEO, Mark Zuckerberg, was interviewed by the US congress in connection with the scandal, he was faced with questions regarding the platform's responsibility when disseminating content. Zuckerberg repeatedly argued that AI was a tool that could be used to combat unwanted content such as hate speech, fake news, revenge porn, etc. His responses have been criticised for expressing a simplistic *"AI solutionism"* – in line with Evgeny Morozov's critical account on *"technological solutionism"*, that is, a sort of coded social engineering based in a firm belief in technology's abilities to solve complex social issues[49] – and for the fact that auto-

45 When the persons running The Pirate Bay file-sharing site were prosecuted in 2009 for complicity in violation of the Copyright Act, a similar conceptual challenge emerged when the court was forced to assess this "platform's" liability; Stefan Larsson, "Metaphors, Law and Digital Phenomena: The Swedish Pirate Bay Court Case", *International Journal of Law and Information Technology*, 21(4), 2013, p. 329-353; Larsson, *Conceptions in the Code. How Metaphors Explain Legal Challenges in Digital Times*. Oxford University Press, 2017a.

46 Cf. Gillespie, 2018.

47 Ulrich Dolata. *Apple, Amazon, Google, Facebook, Microsoft: Market concentration-competition-innovation strategies* (No. 2017-01). Stuttgarter Beiträge zur Organisations-und Innovationsforschung, SOI Discussion Paper, 2017.

48 A news story that received much attention when journalist Carole Cadwalladr published an article about a whistle-blower in The Guardian, 18 March 2018. https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump

49 Evgeny Morozov. *To save everything, click here: The folly of technological solutionism*. Public Affairs, 2013.

mated optimisation tools on which the large-scale platform is based have, in actual fact, contributed to disseminating fake news and controversial content.[50] A responsibly designed platform is faced with a number of normative challenges, such as defining what kind of images, texts and links could be deemed as offensive, unlawful or fake. Often, these are defined differently depending on culture and jurisdiction. Some areas of knowledge, e.g., historical events or geographic definition of regions, can also be controversial and be contested by one of the involved groups, which makes the normative task as complex as it is necessary.

Autonomous Vehicles

A number of traditional car manufacturers around the world are currently developing autonomous vehicles and are facing challenges from technology corporations such as Google's spin-off company Waymo, transport provider Uber and electric car manufacturer Tesla. Public transport company Nobina, based in Kista, Sweden, has conducted unmanned bus tests, and a bus route has been running since 2018. Developers in China, Poland, Switzerland, USA, among other places, are conducting similar, ongoing projects using self-driving public transport vehicles, and it is only a question of time before autonomous vehicles become a common feature of everyday transport in many cities around the world. Automation, which in data-driven applications often largely depends on algorithms designed to perform automation functions, is an area that is of central importance for self-driving vehicles, and raises questions of accountability here too. In Sweden, for example, regulations are being created that address developments in the field of self-driving vehicles[51], and the question of accountability is a key issue in the context of traffic accidents and has also been discussed in the literature for some time.[52] These questions have been raised

50 Kirsten Gollatz, Felix Beer and Christian Katzenbach. "The turn to artificial intelligence in governing communication online," *Social Science Open Access Repository*, 21. 2018. Cf. BuzzFeed News. "Why Facebook Will Never Fully Solve Its Problems With AI", by Davey Alba, 11 April 2018. https://www.buzzfeednews.com/article/daveyalba/mark-zuckerberg-artificial-intelligence-facebook-content-pro

51 Cf. SOU 2018:16, in which delegation of responsibility and data protection issues is a key component.

52 Cf. Alexander Hevelke & Julian Nida-Rümelin. "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis". *Science and Engineering Ethics* 21(3). 2015, p. 619–630.

not least in connection with fatal accidents involving autonomous vehicles. In 2016, a Tesla S model, which uses both radar and cameras to interpret its surroundings, mistook a lorry for the sky, resulting in a fatal accident. In March 2018, a SUV used by Uber to develop self-driving vehicles struck and killed a woman in Arizona, which led to extensive discussions on accountability issues and the use of self-driving vehicles on public roads. Even if comparisons to manned vehicles would show that autonomous vehicles are safer, accidents like this will have an impact on people's trust and acceptance of highly autonomous vehicles.

The Black Box and Algorithmic Transparency

The absence of transparency in connection with algorithm-driven processes, sometimes referred to as "black-boxing", is a well-known problem.[53] Problems related to the delegation of responsibility often have to do with understanding the actual preceding events, even if increased transparency does not solve all problems.[54] Lack of transparency is often described in terms of a trust deficiency, e.g., the EU commission's communiqué on artificial intelligence.[55] The EU Commission is conducting a study in 2018 and 2019 that analyses so-called *algorithmic transparency* for raising awareness and building a good knowledge base for challenges and opportunities for algorithmic decisions, 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."[56] There is a field of studies within AI

53 Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, Franco Giannotti, "A survey of methods for explaining black box models". *ACM Computing Surveys* (CSUR), 51(5), 2018, p. 1-45; cf. Frank Pasquale. *The Black Box Society. The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015;

54 Mike Ananny & Kate Crawford. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability". *New Media & Society*, *20*(3), 2018, p. 973-989.

55 Communication from the commission to the European Parliament, the European Council, the European Economic and Social Committee and the Committee of the Regions. *Artificial Intelligence for Europe* {SWD(2018) 137 final}.

56 EU Commission. Algorithmic Awareness-Building. 25 April 2018. https://ec.europa.eu/digital-single-market/en/algorithmic-awareness-building

research that focuses on the *explainability* of algorithmic complex processes (see point 7 below).

Here I suggest an additional six nuances or aspects of transparency to take into account for the analysis of applied AI on markets, as aspects of AI governance. A challenge, from a societal and legal perspective, lies in balancing opposing interests, where points 1 and 2 below represent counteracting interests and 3 to 7 constitute variants of knowledge and other transparency challenges.

1. Proprietorship

A proprietary approach with corporate software and data is a legitimate way of conducting competitive innovation with a commercial logic. It can be the result of commercialization and upscaling of a product, and can constitute a prerequisite for investors. Some companies view the user data they hold as being directly related to their stock market value, and their software and algorithms as valuable "recipes" and business secrets.[57] However, proprietary set-ups involving company-owned software and data are often referenced as a problematic issue in discussions on overview and scrutiny practices.[58] At worst, and according to Rashida Richardson of the AI Now Institute, proprietary set-ups may "inhibit necessary government oversight and enforcement of consumer protection laws" in that it contributes to the black box effect.[59] This may be particularly problematic for public sector procurement. For example, one component of the challenge posed by the aforementioned COMPAS example regarding the risks of recidivism is the lack of transparency and ensuing lack of informative feedback.[60]

57 Sarah Spiekermann & Jana Korunovska, "Towards a value theory for personal data". *Journal of Information Technology*, 23(1). 2016, p. 62-84. doi:10.1057/jit.2016.4

58 Cf. Pasquale, 2015.

59 Rashida Richardson, "Optimizing for Engagement: Understanding the Use of Persuasive Technology on Internet Platforms". AI Now Institute: *statement before the United States Senate Committee on Commerce, Science, and Transportation. Subcommittee on Communications, Technology, Innovation and the Internet*. June 25, 2019, p. 6.

60 Cf. Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Allen Lane, 2016.

2. Avoiding Abuse

Some algorithm-dependent and automated processes could be abused if the affected parties were made aware of their precise functions. Transparency can, at worst, lead to manipulation or *gaming* of the purpose of a process. This could apply for various types of processes guided by AI where there is an incentive to manipulate the results; such as search engines, trending topics in Twitter[61], welfare distribution, fraud detection practices used by both insurance companies and banks; and even organ matching.

3. Literacy

For the everyday dispersion of new technologies, here applied AI, the *data literacy* or *algorithm literacy* can be one additionally fruitful way to conceptualize how individual's abilities interact with the technologies, implicating their transparency.[62] To even begin to assess algorithms and how they use data, specific expertise is required that people in general do not have. The importance of this type of literacy can also be expanded to an argument targeting contemporary supervisory authorities that are increasingly struggling with supervising data-driven and automated markets and activities (see also point 6 below).[63]

4. Concepts, Terminology and Metaphor

The language, metaphors and symbolism inherent in explanations of complex AI processes have a direct impact on how they are understood. Explanations, however, can be phrased differently depending on the required level of explainability and inherent symbolism, or social need,[64] which complicates matters when analysing how to formulate explanations

61 Caplan et al., 2018, point out that only the slightest disclosure of how Twitter's *trending* method works has made it possible to manipulate parts of their environment and fill selected topics with automated bots or bot-networks in order to influence, manipulate or simply ruin discussions.

62 Derived from media and information literacy, cf. Jutta Haider, & Olof Sundin. *Invisible Search and Online Search Engines: The ubiquity of search in everyday life*. Chicago: Routledge Studies in Library and Information Science, 2019.

63 Larsson, 2018.

64 Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, & Alexandra Wood. "Accountability of AI under the law: The role of explanation." *arXiv preprint arXiv:1711.01134*, 2017.

(see also point 7 below). For example, when formulating an explanation of how AI-generated decision-making works, a decision must unavoidably be made regarding what symbols or metaphors are appropriate at different levels of concretion. I have elsewhere shown that the metaphors used to explain complex digital phenomena will have an effect on normative and legal positions. This has partly to do with historical conditions, i.e., earlier conceptual path dependencies that influence how we understand things by framing them in terms of previously established concepts.[65] The metaphors and symbolism used to explain AI-generated processes will therefore likely have a strong impact on how they are understood or accepted.

5. Complex Data Ecosystems

The lack of transparency can be related to how contemporary AI very much depends on access to large amounts of data, that is collected, traded and brokered on global information markets that can be labelled as "ecosystems". These consist of a number of actors and data brokers, which is, for example, evident in the complexity of this matter.[66] Pasquale states that it is unreasonable for data brokers to presume that individuals will claim their data protection rights in all dealings with every single data-broker.[67] For example, the real-time bidding (RTB) in adtech markets have been stated to be particularly opaque and complex (and lacking consent) in its automated setup with a large number of involved actors.[68]

6. Distributed, Personalised Outcomes

Relevant, personalised services, such as Google's search engine, targeted marketing, or Facebook's personalised news feeds, lead to highly distributed outcomes. From a transparency perspective, the challenge of distributed and personalised outcomes lies primarily in the difficulties

65 Larsson, 2017a.

66 Wolfie Christl. *Corporate Surveillance in Everyday Life: How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions*. Vienna: Cracked Labs, 2017.

67 Frank Pasquale. "Exploring the Fintech Landscape." Written Testimony of Frank Pasquale Before the United States Senate Committee on the Banking, Housing, and Urban Affairs. 2017, September 12; Stefan Larsson. "Algorithmic Governance and the Need for Consumer Empowerment in Data-driven Markets", *Internet Policy Review* 7(2). 2018, p. 1-12.

68 Information Commissioner's Office (ICO), UK. *Update Report into Adtech and Real Time Bidding.* 20 June 2019.

of discovering inappropriate patterns in actions that are only apparent in personalised, sometimes deeply private, matters. Enforcement efforts by supervisory authorities can be seen as an attempt to increase transparency to gain a better overview of these providers' services in order to thereafter assess whether any practices can be deemed improper. In an article on consumer protection rights in the context of data-driven and automated industries, e.g., online marketing in social networks, I argue for the need for *algorithmic governance,* in terms of that supervisory authorities need to improve their methods if they are to discover structural irregularities or illegal outcomes derived from automated AI-driven systems.[69]

7. XAI and Algorithm Complexity

As mentioned, there is an inherent problem in assessing individual outcomes of complex AI tools. Within the area of AI research, a specific field (XAI) that deals with explainability or interpretability has emerged in response to problems related to machine learning, which also entails a "black box" for researchers: i.e., a problem may be sufficiently solved, but it is not possible to precisely interpret how it was solved. The results may indicate a higher probability of a certain outcome, e.g., it may lead to improved profitability or more precise predictions, but not necessarily to a more detailed understanding of how the results were achieved. A critical review shows the need to classify the problems more clearly,[70] not least in relation to the increased practical significance,[71] and where knowledge in social scientific disciplines such as social psychology and cognitive science could also contribute.[72]

69 Larsson, 2018.

70 Guidotti et al., 2018.

71 Biran & Cotton, 2017.

72 Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence*. 2019, Vol 267: 1-38. https://doi.org/10.1016/j.artint.2018.07.007.

Discussion: Mirrors and Norms

THE BASIC TENETS of justice have been a key in general jurisprudential literature throughout the years, and will be a source for further dispute and a recurring point of discussions on the implications of artificial intelligence. Hildebrandt argues that a number of fundamental rights are at risk in a society that is managed using data-driven agency and smart technologies.[73] Analysing the relation between morality and law, not least in the context of justice, has been a key issue for many early legal theorists, for example the Polish legal sociologist Leon Petrazycki, who wrote the body of his work in St Petersburg and Warsaw in the early 1900s. Petrazycki distinguishes, for example, between positive and intuitive law as well as official and unofficial law, the latter being reminiscent of Eugen Ehrlich's concept of a "living" law that is reproduced informally in society.[74] In doing so, he allowed for a more empirically based approach to law which has greatly influenced many later researchers. This informal, contextual, and possibly fluid notion of norms may help us understand that artificial intelligence not only has the capacity to imitate behaviours and linguistic conventions but also has the potential to learn from social norms in order to act as an autonomous agent in possession of normative agency. It will in this process have to *choose* which norms to learn from,[75] opening up for conflict between different sets of informal norms, or conflict between social and legal norms.[76] This could for example regard different

73 HILDEBRANDT, 2015, p. 133ff.
74 EHRLICH, 2002.
75 Cf. IEEE, 2019, p. 36.
76 Cf. SVENSSON & LARSSON, 2012.

groups, ethnicities, religions, demographics with different notions of what is regarded as right and wrong for everything from families, nudity, gender, sexuality, to free speech, media habits, driving behaviour, and so on. This is particularly evident for content moderation in social media platforms, as indicated above.[77] Choosing which norms to learn from, may be a key challenge as AI engages and interacts with human social structures. In addition, as the systems gain in agency, a key question would be to address what responsibility the developer of autonomous agents has for the contents produced by the agents.

The Mirror Effect: Accountability for Reproducing Social Bias

One unavoidable question on the topic of developers of services that learn from inherent, structural values and social conditions concerns how to deal with social bias: should they reproduce the world in its current state or as we would prefer the world to be? And who gets to decide which future is more desirable?[78] Data-dependent AI that learns from real world examples derived from human activities may be understood as a mirror for social structures, leading to questions of accountability for those devising the mirror, its reproducing as well as amplifying abilities. Potentially, there are a number of algorithm-dependent situations in which said algorithms lead to not only automated but normative decisions. It is important to realise that applications that use data retrieved from social contexts not only may produce beneficially "personalized" and individually relevant products and services, but also may contain a number of structural biases and imbalances that societies struggle with in general, such as inequality, unfairness, discrimination and racism. These may lead to normative questions for the designing side, that is, the platforms or data-driven applications that utilise and automate self-learning technologies will ultimately face the normative question of what

77 Cf. Gillespie, 2018.

78 E.g., as noted by researchers and published in Nature; James Zou & Londa Schiebinger. "AI can be sexist and racist – it's time to make it fair", *Nature*, comment, 18 July 2018.

the application *ought to* reproduce or not. And, consequently, be held accountable for the agency it thereby represents as it interacts with and reproduces a biased society. Conversely, this means that AI-driven analytical methods may reveal biases in already present and historical decision-making, which at best can be used as a tool for detection, which also may come as an unpleasant surprise in some cases.

There is an increasing awareness, as noted for example in the aforementioned IEEE report and in several reports published by the AI Now research centre, that cultural values and social biases are inherent components of personal data and must therefore be managed responsibly in software design.[79] However, from a socio-legal perspective, it can be concluded that there are rarely simple solutions or "quick fixes" when addressing normative issues, particularly not for the scale of digital platforms operating with multiple billions of users globally. For want of a truly neutral stance, AI developers will have to adopt normative positions on issues they probably would prefer to avoid, which lends weight to the argument that programs for training AI engineers in image analyses and algorithms should also address the issue of accountability and social or ethical consequences of the designs they are taught to implement and develop.[80] It is also conceivable that this should be addressed in board meetings of companies that operate in consumer markets. Naturally, the primary objective of said companies is to increase revenue, e.g., by way of increasing accuracy in targeted marketing or personalised services, but at what cost and in accordance with what ethical considerations? For example, may personalised pricing by proxy potentially lead to so-called technological redlining? Can automated analytical methods unintendedly lead to a manipulating rather than a fair influencing of consumers? Consider for example "hypernudging", that is, what can be called automated and predictive data-driven decision-guidance techniques.[81]

79 Cf. Whittaker et al., 2018.

80 Cf. Whittaker et al., 2018, p. 6, point 10.

81 Karen Yeung "'Hypernudge': Big Data as a mode of regulation by design", *Information, Communication & Society, 20:1,* 2017, p. 118–136.

Normativity in design, in this context, is a crucial issue. For many AI applications, particularly those that interact with human values and social structures, there is arguably no truly neutral position to find since different situations may require controversial, normative decisions. An image database that has a gender bias might, for example, be descriptively correct in that it might describe contemporary, unequal social conditions in which women are predominantly portrayed in kitchen settings while men are portrayed as being out hunting (as in the previous example), or it may base its assessments on unequal income for the same work; further, applications that "learn" from these conditions also become active agents in this unequal environment. Developers could therefore, unwittingly or unwillingly, end up in a normative position on whether they ought to reinforce or counteract such conditions.

Conclusions: Socio-Legal AI Studies

THE GOAL OF the present text has been to contribute to a broad socio-legal orientation by describing some of the legal and normative challenges of AI. I have drawn on socio-legal theory in relation to growing concerns of fairness, accountability and transparency of applied AI and machine learning in society, to stress the need for AI research and development to keep society "in-the-loop" by utilising insights from fields such as law and society.[82] Specifically, the argument has been focusing normativity in design, societal bias in autonomous and algorithmic systems, as well as difficulties with distribution of liability and accountability, particularly in relation to issues of transparency.

The argument that designing AI is a normative process recognizes that knowledge of cultural values, norms and ethics must, in that case, be implemented in AI developments and applications in order to be able to address aforementioned risks. Since AI and machine learning, when appropriately implemented, have indisputable potential social benefits, it could be said that the social perspective implies a need to understand how we should proceed to achieve trust and social acceptance in these applications.[83] We can therefore conclude that an appropriate level of transparency, well thought-out delegation of algorithmic accountability and clear indications that autonomous systems do not strengthen

82 RAHWAN, 2018.

83 This is in line with for example AI HLEG's Ethics guidelines for trustworthy AI (2019); the IEEE's Ethically Aligned Design, 2019; and Luciano Floridi, et al. "AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations". *Minds and Machines,* 28, 2018, p. 689–707.

or reproduce social biases and prejudices in an unjust manner, or in any other way are detrimental to basic social functions, are crucial for establishing trust in the system.

In discussions on regulation – whether they revolve around the need for new regulations, or laws that lag behind, or digital platform companies arguing for self-regulation in a technological solutionist manner – it should be remembered that well-established regulations that have broad legitimacy already exist for many aspects and applications which use data-driven artificial intelligence. Grounds for addressing discriminatory practices, market laws, and data protection regulations already exist. The challenges that face these kinds of regulations, in the context of autonomous systems, often have to do with how to discover problems, regulate and implement solutions, but also, how to address the conceptual issue of translating conventional views on discrimination, co-determination and unfair practices to new market practices.

The most important conclusions are:

1. *The need for an interdisciplinary and multidisciplinary approach:* A crucial insight from recent research on FAT and working groups on ethical guidelines for AI is that the combination of AI and society demands multidisciplinary research to be responsibly developed into trusted applications. Contemporary data-dependent AI should not be developed in a technological isolation without continuous assessments from the perspective of ethics, cultures and law. This can be exemplified by the multidisciplinary approach on the challenges of AI transparency described above. It means that we need to increase our awareness in matters concerning values and normativity, as well as multidisciplinary and interdisciplinary approaches to research, development and education. Neither should fields that address ethical, legal and social issues be seen as a superficial layer overlying current AI developments in computer science or mathematical institutions, but rather, as important, complementary fields of expertise that can contribute to AI research, algorithm developments and machine

learning. Some applications have become notorious as a result of bad design caused by an exaggerated reliance on one-sided skillsets.

2. *Principles without processes are ineffectual:* Albeit much effort laudably is put into producing principles to govern applied AI, recognizing that normativity is an important aspect also necessarily entails implementing some form of process. There are lessons to be learned from centuries of developing legal orders and legal processes when it comes to establishing and implementing principles for AI and machine learning; e.g., comparisons can be made to how prosecution procedures need to comply with norms; comparisons between how the various supervisory powers and judicial power are organized; how general principles can be related to individual cases, etc.
3. *The importance of context:* Recognising normativity as an empirical phenomenon unavoidably entails encountering and dealing with contextual deviations and blatant normative contradictions: which norms should apply? For example, as large scale digital platforms gain billions of active users they inevitably operate in a large number of cultures, communities and jurisdictions consisting of different cultural preferences, and possibly contradictory takes on a number of issues relating to family norms, sexuality and relationships, nudity, ethnicity and social status, etc.
4. *The need for supervisory competence and impact assessment:* It is necessary to develop methods for supervisory authorities in light of the fact that automated AI and machine learning have the potential to provide highly decentralised outcomes in which transparency is primarily afforded to individual users or addressees. Methods are needed to discover discriminatory patterns or other improper practices at a structural level, such as the aforementioned "redlining" issue, as well as to standardise societal impact assessments of AI processes in relation to consumer markets and the public sector.
5. *The balancing of transparency:* Arguably, while one of the core challenges with applied AI is dealing with explainability and opaqueness of so-called black box applications, AI transparency opens for a complex set of interests to be balanced. The benefits of each kind of application need to be weighted at a societal level to determine the most

appropriate degree of transparency. The importance of transparency and explainability needs to be assessed in relation to stakes and needs posed in each context, which may mean that translations to ethical and legal needs will be required.

It is important to emphasise that a focus on these challenges should not discourage efforts to apply a normative perspective to artificial intelligence. Rather, the intent is to contribute to, and clarify, issues that need to be developed further and require greater knowledge and awareness. To a large degree, we already live in a highly digitalised environment in which the data we generate in our daily lives is increasingly used and reused as training data for self-learning technologies in automated processes and autonomous decision-making. There are strong indications that our lives will increasingly be enabled and affected by different kinds of artificial intelligence and machine learning in the years to come, since these methods and technologies have already been proven to have great potential. This means that it becomes all the more important to strengthen fairness and trust in applied AI through well-advised notions of accountability and transparency in multidisciplinary research of socio-legal relevance.

Acknowledgements

I would like to extend my thanks to the International Institute of the Sociology of Law in Oñati, the Basque Country, for my research stay in June and July 2018, and for allowing me to use their well-stocked library while preparing an early draft of this article.

THIS REPORT DRAWS on socio-legal theory in relation to growing concerns over fairness, accountability and transparency of societally applied artificial intelligence (AI) and machine learning. The purpose is to contribute to a broad socio-legal orientation by describing legal and normative challenges posed by applied AI. To do so, the report first analyses a set of problematic cases, e.g. image recognition based on gender-biased data- bases. It then presents seven aspects of transparency that may complement notions of explainable AI within computer scientific AI-research. The report finally discusses the normative mirroring effect of using human values and societal structures as training data for learning technologies, and concludes by arguing for the need for a multidisciplinary approach in AI research, development and governance.

This report is a preprint of Stefan Larsson's article The Socio-Legal Relevance of Artificial Intelligence in forthcoming special issue "Le droit à l'épreuve des algorithmes" in Droit et société, 103(3), which is edited by Dubois & Schoenaers.