www.jusletter.ch

Anne-Sophie Morand

A WEIRD AI system?

Inequalities and power asymmetries in the context of Health Tech

Entwicklungen in der Health-Tech-Branche bieten eine Vielzahl von Möglichkeiten und Chancen. So kann der Einsatz von künstlicher Intelligenz beispielsweise zu einer Verbesserung der Heilungschancen und damit zu einer Senkung der Gesundheitskosten führen. Allerdings können sog. «Automated Decision-Making Systems» auch Verzerrungen und Diskriminierungen widerspiegeln. Die Autorin geht dieser Problematik nicht nur aus technischer Sicht auf den Grund, sondern beschreibt auch, wie die aktuellen Pläne für eine Regulierung von KI in Europa aussehen, damit künftig «faire Algorithmen» Entscheidungen treffen und Machtasymmetrien verhindert werden.

Category of articles: Articles
Field of Law: Health law

Citation: Anne-Sophie Morand, A WEIRD AI system?, in: Jusletter 20 September 2021

ISSN 1424-7410, jusletter.weblaw.ch, Weblaw AG, info@weblaw.ch, T +41 31 380 57 77

Contents

1. Introduction

[1] Medical services are more and more moving away from the classic doctor-patient relationship. Technical innovations are increasingly replacing the relationship of trust. An essential factor of digital services in the health tech sector are algorithms that support or even replace doctors. For example, software that supports doctors in making diagnoses or evaluating MRI results, but also medical applications that can be downloaded by anyone to a smartphone and used without the involvement of a doctor. For instance, the risk assessment of a skin cancer app does not come from a doctor but an algorithm.[1]

[2] Developments in the health tech industry can lead to an improvement in patients' chances of recovery and thus to a reduction in healthcare costs.[2] For example, medical apps such as the aforementioned skin cancer app promote early detection and therefore improve the health care system, especially among people who avoid doctor's visits and in areas with a low number of doctors.[3]

1 Such apps already exist (e.g. https://www.skinvision.com/au/); all websites in this article last visited on 19 July 2021).

2 PwC, Sherlock in Health – How artificial intelligence may improve quality and efficiency, whilst reducing healthcare costs in Europe, June 2017, p. 1 et seqq. (https://www.pwc.nl/nl/assets/documents/pwc-sherlock-in-health.pdf; cit. PwC study); The opportunity costs of not using the best available tools for disease detection and treatment are substantial – millions of people receive misdiagnoses every year. For example, nearly one third of all preventable deaths in the United Kingdom in 2018 were attributable to misdiagnosis. The benefits of early disease detection through AI are immense; *see* David Watson/Jenny Krutzinna/Ian Bruce and Co., Clinical applications of machine learning algorithms: beyond the black box, in: BMJ 2019, p. 364 et seqq.

3 Corinne Widmer Lüchinger, Apps, Algorithmen und Roboter in der Medizin: Haftungsrechtliche Herausforderungen, HAVE, p. 4 (cit. Widmer Lüchinger).

[3] Despite all the positive aspects, the question arises whether Artificial Intelligence (AI) health support promises the same success for everyone. Cases are discovered in which algorithmically controlled, Automated Decision-Making (ADM) systems make decisions to the detriment of members of minorities and vulnerable groups. That is alarming, especially when these disadvantageous decisions reflect and perpetuate social inequalities. The problem can arise, for example, when the algorithmic system fails to detect skin cancer amongst dark-skinned people.

[4] In a first step of this article, important terms such as «algorithms» are defined and explained. Based on this knowledge, the author explains how AI is used in the health tech sector. Special attention is then paid to the discriminations and inequalities that can arise through ADM systems in the health sector. The author uses various examples to show what it means for the equal treatment of people when important decisions in the health sector are delegated to such systems. Secondly, it makes not only sense to analyse the extent to which AI can entrench and deepen existing inequalities and power asymmetries by using the example of health tech but also, to take a closer look at how this problem is addressed from a legal point of view. The European Union has recognised the problem of existing inequalities, which is one reason why the EU Commission presented its first regulatory proposal on AI on 21 April 2021. This regulation – similar to the General Data Protection Regulation (GDPR) – is likely to become of international importance. In this case, it would also have an impact on Switzerland. This article thus shortly discusses current legal developments in the EU and in Switzerland.

2. Algorithmic Decision-Making Systems

[5] ADM systems are procedures in which decisions are initially delegated to another person or corporate entity, who use data-driven, automatically executed decision-making models to take an action and present a decision. ADM systems are based on algorithms.

2.1. What is an Algorithm?

[6] Algorithms can be defined as «*a set of rules defining how to perform a task or solve a problem.*»[4] They consist of instructions such as «if this happens, then that is to be done». Generally speaking, algorithms are used as a term for programmed procedures that calculate a specific output from a specific input by means of a precisely defined, serial sequence of steps.[5] Ideally, the instructions are formulated in such a way that no room for interpretation remains open. They can calculate both less complex outputs, such as sorting a series of numbers according to their size, and highly

4 European Parliamentary Research Service (EPRS) Study, Artificial intelligence: How does it work, why does it matter, and what can we do about it?, 2020, p. VI (https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641547/EPRS_STU(2020)641547_EN.pdf; cit. EPRS Study); Paul von Bünau, Künstliche Intelligenz im Recht, 2018, p. 80 (https://legal-revolution.com/images/pdf/KI-Im-Recht.pdf; cit. von Bünau).

5 Kilian Vieth/Ben Wagner, Wie algorithmische Prozesse Teilhabechancen beeinflussen können, im Auftrag der Bertelsmann Stiftung, 2017, p. 9 (https://algorithmenethik.de/wp-content/uploads/sites/10/2018/02/Teilhabe-ausgerechnet.pdf).

complex outputs, such as detecting cancer by analysing photos of the skin. An algorithm is, therefore, a sequence of well-defined, step-by-step instructions to solve problems.[6]

2.2. What is Artificial Intelligence?

[7] Intelligence demonstrated through the use of algorithms is referred to as AI, a subfield of computer science. AI systems demonstrate characteristics that we associate with intelligence from humans, namely language comprehension, learning, reasoning and problem solving.[7]

2.2.1. The term «AI»

[8] The beginnings of a definition of AI can be found in the literature of John McCarthy, one of the first experts in the field of AI. He defines AI as «*science and engineering of making intelligent machines, especially intelligent computer program.*»[8] Concerning the term «Intelligence», he states: «*Intelligence is the computational part of the ability to achieve goals in the world. Varying kinds and degrees of intelligence occur in people, many animals and some machines*».[9] McCarthy claims that «it does not exist a solid definition of intelligence that does not depend on relating it to human intelligence».[10] In science, it is still not clearly established how «Intelligence» can be defined and what characteristics an AI must have. Today's AI is known as «weak AI» – these AI systems are powerful in particular and specific tasks.[11] «Weak AI» is characterised by pattern recognition and the ability to react to unknown problems. However, weak AI cannot abstract and can only be used in a specific field of application. Despite the classification of today's AI applications as «weak AI» they're nevertheless capable of high technical performance. However, «strong AI» – which does not exist yet – could flexibly adjust to all sorts of tasks a human being could do.[12] It can be concluded that AI deals with methods that make it possible to simulate human intelligence as closely as possible on computer systems with the help of ADM systems.

[9] The EU's new draft law on AI regulation, published on 21 April 2021, defines AI systems as softwares which use one or more techniques and approaches that can produce human-determined outcomes affecting the environment they interact with. In concrete, the future-oriented definition is as follows: «*Artificial intelligence is a fast-evolving family of technologies that can contribute to a wide array of economic and societal benefits across the entire spectrum of industries and social activities. By improving prediction, optimising operations and resource allocation, and personalising digital solutions available for individuals and organisations, the use of artificial intelligence can provide key*

6 Florian Saurwein, Automatisierung, Algorithmen, Accountability. Eine Governance Perspektive, in: Matthias Rath/Friedrich Krotz/Matthias Karmasin (Hrsg.), Maschinenethik. Normative Grenzen autonomer Systeme, 2019, p. 35.

7 Thomas Söbbing, Künstliche Intelligenz im HR-Recruiting-Prozess: Rechtliche Rahmenbedingungen und Möglichkeiten, in: Torsten Kutschke/Stefan Müller (Hrsg.), Zeitschrift zum Innovations- und Technikrecht, p. 64 (cit. Söbbing); EPRS Study (Fn. 4), p. 1.

8 John McCarthy, What is artificial intelligence?, 2007, p. 2 (http://jmc.stanford.edu/articles/whatisai/whatisai.pdf; cit. McCarthy).

9 *Ibid*; Söbbing (Fn. 7), p. 64.

10 McCarthy (Fn. 8), p. 3.

11 Söbbing (Fn. 7), p. 65.

12 Dirk Helbing, What's Wrong with AI? A Discussion Paper, 2020 (https://magazine.swissinformatics.org/de/whats-wrong-with-ai/).

competitive advantages to companies and support socially and environmentally beneficial outcomes, for example in healthcare, farming, education and training, infrastructure management, energy, transport and logistics, public services, security, justice, resource and energy efficiency, and climate change mitigation and adaptation».[13] This definition of AI systems is fairly broad, which on the one hand, allows flexibility with regard to the fast-paced technological developments in AI systems, but on the other hand, can also cause legal uncertainty for developers, operators and users of AI systems.

2.2.2. Machine Learning

[10] Machine Learning (ML) is a form of AI, using algorithms, and enables systems to learn autonomously from data, to recognise complex relationships within large data sets and to improve themselves without the need for explicit programming of each individual step.[14] ML is based on the idea of gaining knowledge from experience.[15] In other words, ML algorithms develop understanding, make decisions, and evaluate their confidence from the training data. For this purpose, a machine receives concrete sample data with known contexts from which a general rule is to be extracted in order to later apply this structure to unknown contexts. A dataset has both rows and columns, with each row containing one observation. This observation can be, for example, an image, a text or a video.[16] The quality of the sample data is important and finally has a huge impact on the effectiveness and efficiency of the algorithm. The better and fairer the training data is, the better and fairer the model can perform.

[11] When explaining ML, it is important to consider the «problem of correlation without causation». This problem occurs when the algorithm is mistaking correlation with causation. The following illustrative example will serve to illustrate the problem: 100 patients are hospitalised with pulmonary infections. 15 of them also have asthma and it is known that asthma increases their risk of getting sicker. Therefore, the doctors give more aggressive treatment to the patients with asthma. As a result, the patients with asthma actually recover faster and better. Assuming this data is used to train an ML model, the model might conclude that patients with asthma have a better recovery process. As a result, the AI could recommend that asthma patients should be treated less aggressively, when in fact the opposite would be accurate. Hence, it can be stated that ML is able to identify complex relationships within large volumes of data to predict outcomes with high accuracy. The problem, however, is that these relationships are often just correlations and not causalities.[17]

---

13 Proposal for a Regulation of the European Parliament and of the Council of 21 April 2021, Laying down harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, 2021/0106(COD), Art. 3 (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN; cit. Proposal AI Act).

14 Adewole Adamson, Machine Learning and Health Care Disparities in Dermatology, November 2018, p. 1247 (https://jamanetwork.com/journals/jamadermatology/article-abstract/2688587; cit. Adamson).

15 Alex Mitchell/Sanjay Rao/Amol Vaze, Clinical diagnosis of depression in primary care: a meta-analysis, in: Lancet 374, 2009, p. 609 et seqq.

16 Adamson (Fn. 14), p. 1247; von Bünau (Fn. 4), p. 81; Markus von Rimscha, Algorithmen kompakt und verständlich, Lösungsstrategien am Computer, 2017, p. 132 (cit. von Rimscha).

17 Chris Lovejoy, From Correlation to Causation in Machine Learning: Why and How, 31 October 2020 (https://towardsdatascience.com/from-correlation-to-causation-in-machine-learning-why-and-how-4485bca8d145).

2.2.3. Deep Learning

[12] Deep Learning (DL) is a subset of ML. It imitates human learning behaviour in processing large amounts of data. For this purpose, learning processes are implemented in artificial, multi-layer neural networks which are capable of learning unsupervised from data that is unstructured or unlabeled. The neural networks consist of neurons that are similar to the synapses of humans.[18] DL uses different layers in the neural networks:

1. The first layer of the neural network, the visible input layer, processes a raw data input, for example the single pixels of an image (*Input Layer*).
2. The information is further processed and reduced via several hidden layers (*Hidden Layer*).
3. The output layer finally leads to the result (*Output Layer*).[19]

[13] The input layer is linked to the output layer in different ways via the hidden layer. In retrospect, however, it is usually impossible to understand which decisions were made on the basis of which data – the machine automatically refines the decision rules.[20] DL refers to models that have more than one hidden layer. Thus, neural networks are «able to learn» and can, for example, distinguish between different objects in an image by following rules defined by the network itself. DL systems are capable of recognising the smallest dependencies between variables and – in contrast to classical ML – transform unstructured data into numerical values.[21]

3. Artificial Intelligence Systems in the Health Sector

[14] Health technologies support people in the health sector with tools that are important for effective and efficient prevention, diagnosis, treatment and rehabilitation and the achievement of internationally agreed health-related development goals.[22] One of the areas in health where AI is the most successful is in diagnostics.[23] When analysing patient data, DL algorithms can rapidly analyse large amounts of information. AI then makes suggestions for action, e.g. recommending a therapy. Moreover, AI generates medication plans, delivers early warnings about the development of chronic diseases or prepares prescriptions for patients.[24] For instance, *Watson*, a computer program developed by IBM, uses algorithms to help doctors to diagnose cancer.[25] The software's findings are at least as precise as those of a doctor, but the algorithm is much faster.[26] Also interesting is the decision support pool *EchoGo Pro*, an AI service for automated identifi-

---

18 Laurzen Wuttke, Machine Learning vs. Deep Learning: Wo ist der Unterschied (https://datasolut.com/machine-learning-vs-deep-learning/; cit. Wuttke); von Rimscha (Fn. 16), p. 158.

19 EPRS Study (Fn. 4), p. 4.

20 Wuttke (Fn. 18).

21 *Ibid*.

22 Resolution on health technologies, WHO, WHA60.29, Sixtieth world health Assembly, 23 May 2007, p. 106 (https://www.who.int/healthsystems/WHA60_29.pdf?ua=1).

23 Marc Jannes/Minou Friele/Christiane Jannes/Christiane Woopen, Algorithmen in der digitalen Gesundheitsversorgung – Eine interdisziplinäre Analyse, im Auftrag der Bertelsmann Stiftung, 2018, p. 15 https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/ Publikationen/GrauePublikationen/VV_Studie_Algorithmen.pdf.

24 Widmer Lüchinger (Fn. 3), p. 2 et seqq.

25 Watson Health IBM, (https://www.ibm.com/watson-health/solutions/cancer-research-treatment).

26 Widmer Lüchinger (Fn. 3), p. 4.

cation of coronary artery disease. The system achieves a diagnostic performance of over 90% and significantly reduces the number of misdiagnoses compared to reports of the routine clinical practice.[27] The field of image-based diagnostics in particular is developing rapidly. This is no coincidence, because AI can recognise patterns in large amounts of data and thus detect disease signals much more precisely, earlier and faster than the human eye.[28] AI can make statements about diseases not only from images, but also from the voice. For example, MIT scientists have developed an app that can tell whether a person is infected with COVID-19 or not just by the sound of a cough. The researchers found that people who are asymptomatic differ from healthy individuals in the way that they cough. The AI system was trained with tens of thousands of forced-cough recordings. The algorithms accurately identified 98.5% of coughs from people who had COVID-19, and 100% of coughs from asymptomatics who held a positive test result.[29]

[15] Regarding the examples mentioned, it can be said that the use of ADM systems in the health sector is linked to a variety of hopes, e.g. the early detection of diseases and an increase in the efficiency of diagnostics and therapy. However, the use of AI in the health sector can further increase (existing) inequalities between different groups of the population.[30]

4. Inequitable Decisions in the context of Health Tech

4.1. Examples of biased decisions in Health Tech

4.1.1. Inequality based on race, skin colour and financial situation

[16] Outdated or biased algorithms can lead doctors to misdiagnose minorities and black people. For example, the American Heart Association heart failure risk score, which ranges from 0 to 100, adds 3 points for non-blacks because non-black patients are identified as more likely to die from heart disease. Or, a kidney stone algorithm added more points for non-blacks, rating them as more likely to have kidney stones. However, in both cases the assumptions were wrong. AI developers sometimes make similar assumptions when developing their algorithms. Thus, algorithms that take race into account may be based on incorrect generalisations and lead to false inferences. In any case, skin colour alone does not explain differences in health risks or outcomes.[31]

[17] In 2019, a study found that an algorithm in the USA favoured white patients over African Americans in selecting chronically ill patients for a program that cares for high-risk patients. In

27 Ultromics (https://www.ultromics.com/).

28 Widmer Lüchinger (Fn. 3), p. 4; Alexia Sidiropoulos, Haftung für Gerätefehler bei der medizinischen Diagnostik und Behandlung, in: Sicherheit und Recht 1/2020, p. 50.

29 Jordi Laguarta/Ferran Hueto/Brian Subirana, COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings, in: IEEE Open Journal of Engineering in Medicine and Biology, 29 September 2020, p. 275; Jennifer Chu, Artificial intelligence model detects asymptomatic Covid-19 infections through cellphone-recorded coughs Results might provide a convenient screening tool for people who may not suspect they are infected, 29 October 2020 (https://news.mit.edu/2020/covid-19-cough-cellphone-detection-1029).

30 Alma Kolleck/Carsten Orwat, Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick, 2020, p. 45 (https://www.tab-beim-bundestag.de/de/pdf/publikationen/berichte/TAB-Hintergrundpapier-hp024.pdf; cit. Kolleck/Orwat).

31 Sharona Hoffman, Biased AI can be bad for your health – here's how to promote algorithmic fairness, 9 March 2021, *passim* (https://theconversation.com/biased-ai-can-be-bad-for-your-health-heres-how-to-promote-algorithmic-fairness-153088; cit. Hoffman).

concrete, African Americans were less likely to be proposed for extra care than white patients for the same disease severity. By using the AI system, African Americans comprised just 18% of the high-risk group. In fact, the high-risk group should have comprised 47% African Americans. In conclusion, white patients had access to resources ahead of African Americans who were less healthy. The inequality stemmed from the fact that AI used past medical expenditure as an indicator for medical need – historical health costs were used as a proxy for medical need in the AI model, at the same time, however, black patients received fewer health resources due to systemic discrimination in the past and this racial bias was then taken into account. In addition, poverty and also the difficulty of accessing health care often keep African Americans in particular from spending a lot of money on health care compared to white people. Thus, the algorithm misinterpreted the low spending as an indication that African Americans were healthy and deprived them of access to extra care.[32] Finally, such an AI model erroneously concludes that black patients need fewer medical resources than white patients and tends to exacerbate the already existing problem.

[18] Another unequal treatment concerned the software called EPIC, an AI-based tool which helps medical offices identify patients who are likely to miss appointments. The tool enabled doctors to double-book potential no-show visits to avoid loss of income. A primary variable for assessing the likelihood of a no-show was previous missed appointments. As a result, AI disproportionately identified economically disadvantaged people who often had problems with transport, childcare and time off work. However, when they did show up for appointments, the doctors had less time for these patients because of double bookings. The overbooked persons ended up receiving a poorer quality of care. In this case, the potential for explicit discrimination was also evident, as the tool included personal characteristics such as ethnicity, financial class, religion and body mass index, which could lead to health resources being systematically diverted from people who are already marginalised.[33]

[19] Studies have found that the use of AI contributed to further exacerbating the disparate impact of COVID-19 on underrepresented and vulnerable groups in the USA, particularly Black, Asian and other ethnic minorities, older people and people of low socioeconomic status. For example, among African Americans aged 35 to 44, the COVID-19 mortality rate is nine times higher than among white people. Indeed, in the early weeks of the pandemic, there were few COVID-19 testing centres in African American communities in the USA. The decision about who and where to test was guided by AI. The AI system was built from biased data reflecting unfair health systems and was, therefore, itself at high risk of bias – even when sensitive variables such as race were explicitly excluded from the system. The initial focus on wealthy white communities had then the effect that many infections spread quickly through places where people with poor health conditions tended to live. It can be concluded that the demand for rapid technological

32 *Ibid*; Ziad Obermeyer/Brian Powers/Christine Vogeli/Sendhil Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, in: *Science*, Vol. *366* (6464), 25 October 2019, p. 366 et seqq. (https://science.sciencemag.org/content/366/6464/447; cit. Obermeyer/Powers/Vogeli/Mullainathan); Thomas Quinn/Stephan Jacobs/Manisha Senadeera/Vuong Le/Simon Coghlan, The Three Ghosts of Medical AI: Can the Black-Box Present Deliver?, 10 December 2020 (https://arxiv.org/pdf/2012.06000.pdf).

33 Sara Murray/Robert Wachter/Russell Cucina, Discrimination By Artificial Intelligence In A Commercial Electronic Health Record – A Case Study, 31 January 2020 (https://www.healthaffairs.org/do/10.1377/hblog20200128.626576/full/).

interventions and the uncritical use of AI in the fight against the COVID-19 arguably hindered responsible development and not-biased use of AI systems.[34]

[20] In most cases, AI systems beat dermatologists in recognising signs of skin cancer on corresponding images. However, in the USA, dermatologists often diagnosed skin cancer too late, especially for dark-skinned patients. The five-year survival rate of black skin cancer patients is thus 73%, whereas 90% of white patients survived the five years after diagnosis. This is because the database that the researchers used to train their AI contained mainly images of white Americans and Europeans – in other words, it contained *WEIRD* data.[35] *WEIRD* stands for western, educated, industrialised, rich and democratic societies.[36] To conclude, the AI was therefore significantly weaker in detecting skin cancer on dark skin than on light skin. When patients with darker skin are underdiagnosed, this can finally lead to further underrepresentation in the clinical data, which again can reinforce the cycle of exclusion.

4.1.2. Inequality based on gender – the «Male Default»

[21] In the health sector, innovations rely largely on male data. Women have often been excluded from medical studies in the past with the argument that their menstrual hormonal fluctuations produce too many variables. The systematic lack of data on women is referred to as the «gender data gap». Even today, a disproportionate number of men participate in medical studies compared to women. Up to now, 80% of painkillers have only been tested on men, although 70% of those suffering from chronic pain are women. The men's symptoms are thus seen as default. However, a «one size fits all» approach in health care can therefore also be dangerously disadvantageous for women.[37] Women are 60% more likely than men to have an adverse reaction to prescription drugs. If healthcare AI systems continue to work with datasets that contain more male data, women will continue to be disadvantaged. For instance, a lack of physiological indicators of heart attacks in women led to AI systems being 50% more likely to misdiagnose heart attacks in women compared to men. Also brain tumours, which appear as headaches in women, are diagnosed with a delay compared to men.[38]

---

34 David Leslie/Anjali Mazumder/Aidan Peppin/Maria Wolters/Alexa Hagerty, Does «AI» stand for augmenting inequality in the era of covid-19 healthcare?, in: BMJ (Clinical research ed.), 15 March 2021, p. 1 et seqq. (https://www.bmj.com/content/bmj/372/bmj.n304.full.pdf); Kat Jercich, AI bias may worsen COVID-19 health disparities for people of color, 18 August 2020 (https://www.healthcareitnews.com/news/ai-bias-may-worsen-covid-19-health-disparities-people-color); Eliane Röösli/Brian Rice/Tina Hernandez-Boussard, Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19, 17 August 2020, p. 190 et seqq. (https://academic.oup.com/jamia/article/28/1/190/5893483).

35 Adamson (Fn. 14), p. 1247; Obermeyer/Powers/Vogeli/Mullainathan (Fn. 32), p. 366; Helene Epstein, Why the Color of Your Skin Can Affect the Quality of Your Diagnosis, p. 5 (https://www.improvediagnosis.org/wp-content/uploads/2020/05/Why-the-Color-of-Your-Skin-Can-Affect-the-Quality-of-Your-Diagnosis.pdf).

36 Thomas Pollet/Tamsin Saxton, How Diverse Are the Samples Used in the Journals «Evolution & Human Behavior» and «Evolutionary Psychology»?, p. 357–368, in: Evolutionary Psychological Science 5, 12 March 2019, p. 357 (https://link.springer.com/content/pdf/10.1007/s40806-019-00192-2.pdf; cit. Pollet/Saxton).

37 Women's Forum, Addressing health barriers through technology, p. 22 et seqq. (https://www.womens-forum.com/wp-content/uploads/2020/04/HEALTH-19e787a3-2643-4548-a840-2085886370b2.pdf).

38 Caroline Criado Perez, Invisible Women: Exposing Data Bias in a World Designed for Men, 2019, p. 195 et seqq.; Annabelle Painter, How can we prevent gender bias in medical AI technology? 8 March 2020 (https://www.babylonhealth.com/blog/tech/ladies-and-gentlemen-lets-kick-some-bias).

4.2. The Black Box Problem

[22] A major advantage of ADM systems, especially of DL, is the ability to analyse enormous amounts of data and to learn patterns and correlations independently. DL algorithms can independently change the decision structures (weighting of the individual nodes within the neural network and independently rearranged relationships between the neural layers) on the basis of the processed data. In these neural networks, many of these layers remain hidden. Since the knowledge in such systems is distributed over several millions or even billions of nodes of a neural network, their methods do not allow a valid insight into the ways of solving the problem. This refers to the enormous complexity of such systems and of the neural connections and mathematical abstractions that these connections generate. Likewise, it is not possible to draw any conclusions from the calculated results about the concrete processing path. Which rules the neural network follows in order to classify the unknown new findings in the existing decision-making system and which variables the individual nodes classify as significant remains invisible to those outside the system. The results of DL systems are therefore no longer comprehensible and the concrete decision paths can only be reconstructed to a limited extent. In other words, as users we know what the input and output is, but we do not know how the system turned the input into the output. This lack of transparency makes the algorithm a «Black Box».[39] Even if a model makes a correct prediction, it may be impossible to understand why the system came to this result.

[23] These technical challenges can lead to problematic results from a social or legal perspective, especially if groups of people are systematically discriminated[40] based on AI decisions or if the outcome of an analysis in sensitive domains cannot be explained. Thus, depending on the area of application, the impacts are different. For example, the lack of transparency in a recommendation for music titles on Spotify is basically unproblematic. On the other hand, an AI system's incomprehensible prediction of a convicted person's risk of recidivism violates fundamental rights of the person concerned. Similarly, the black box problem can lead to inequities within the existing health system or worsen outcomes for vulnerable patients. One can consider the previously mentioned example of the model that was trained to triage patients with pulmonary infections and assigned a low risk to patients with asthma because the model did not consider the intensity of treatment received by these patients with asthma who were more vulnerable.

4.3. «Garbage In, Garbage Out»

[24] In the same way that offline life discrimination takes place, so does human-created technology. In the course of its lifecycle, AI is fed with data by humans or feeds itself with vast amounts of data, most of which originally comes from human beings. It is difficult to obtain data from the past that is not distorted by human biases. In the data mining research community, the term «*Dirty Data*» is often used to refer to these wrong or missing data.[41] AI systems simulate the

39 Will Knight, The Dark Secret at the Heart of AI, MIT Technology Review (https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/).

40 Such a discrimination can only be detected if it is closely analysed and the system is already in place or several well-balanced data sets are available.

41 Rashida Richardson/Jason Schultz/Kate Crawford, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice, New York University Law Review Online, 13 February

world – often a *WEIRD Samples* world – based on the data they have been trained with. If something is missing or incorrect in the data – for instance, pictures of people with dark skin are not included – then the AI model does not recognise the corresponding data set.[42]

[25] In lots of other cases algorithms can be biased toward or against individuals from social or racial minorities.[43] For example, in 2013 it was revealed that the computer program called COMPAS, which was used by US judges, concluded that black defendants will have higher risks of recidivism than they actually do, while white defendants are predicted to have lower rates than they actually do. This systematic discrimination was caused by the training data of the AI which demonstrated an inherent bias.[44] The AI itself had no sense of justice or injustice and it was not apparent how and on what grounds it decided.[45] Such a system ultimately decides without considering the human and social uncertainty factors, unless these factors were present in the learning data.

[26] Nowadays, more and more people are relying on the growing amount of information available through search engines. Therefore, programmers rely themselves on algorithms to sort «who actually sees what». However, if these algorithms are not transparent to the public, it remains unclear why and how a system decided to select a particular image or advertisement. In the end, this reinforcement of bias between the technology and its users can intensify the formation of stereotypes, and then influence how people perceive the world. Online categorisations reflect real world biases and at the same time perpetuate real world discrimination.[46] For example, patterns of marginalisation of disabled people are embedded in data that shape AI systems. Recent research supports this by showing that social attitudes that classify disability as bad are encoded in AI systems that recognise hate speech and identify negative/positive sentiments in written texts. The researchers found that an AI model for analysing sentiments classifies texts that mention disability as rather negative.[47]

[27] To give another example, researchers from Brazil studied algorithms that rated the attractiveness of women on well-known search engines in 59 countries around the world. They wanted to know from the search engines what «beautiful» and what «ugly» women are. In this context, they collected images and identified stereotypes for female physical attractiveness in images on

---

2019, p. 195 (https://ssrn.com/abstract=3333423; cit. Richardson/Schultz/Crawford); Vivienne Ming, Human insight remains essential to beat the bias of algorithms – Better data can improve AI's ability to spot correlations but will not ensure fairness, in: Financial Times, 4 December 2019 (https://www.ft.com/content/59520726-d0c5-11e9-b018-ca4456540ea6); Pollet/Saxton (Fn. 36), p. 358.

42 Meredith Whittaker/Meryl Alper/Cynthia Bennett/Sara Hendren/Liz Kaziunas/Mara Mills/Meredith Ringel Morris/Joy Rankin/Emily Rogers/Marcel Salas/Sarah Myers West, Disability, Bias, and AI, November 2019, p. 9 (https://wecount-dev.inclusivedesign.ca/wp-content/uploads/2020/06/Disability-bias-AI.pdf; cit. Whittaker et al.).

43 Joanne Goodman, Reworking the gender balance in the AI, IoT industries, 12 May 2018 (https://www.controleng.com/articles/reworking-the-gender-balance-in-the-ai-iot-industries/; cit. Goodman); Richardson/Schultz/Crawford (Fn. 41), p. 203 et seqq.

44 Ellora Thadaney Israni, When an Algorithm Helps Send You to Prison, 26 October 2017 (https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html); Kolleck/Orwat (Fn. 30), p. 50 et seqq.

45 Rainer Kessler/Jutta Sonja Oberlin, Künstliche Intelligenz: Quo Vadis?, in: Armin Fladung, Compliance Berater, p. 92.

46 Erika Hayasaki, Is AI Sexist? In the not-so-distant future, artificial intelligence will be smarter than humans. But as the technology develops, absorbing cultural norms from its creators and the internet, it will also be more racist, sexist, and unfriendly to women, 16 January 2017, *passim* (https://foreignpolicy.com/2017/01/16/women-vs-the-machine/).

47 Whittaker et al. (Fn. 42), p. 7 et seqq.

the internet. In most of the countries examined, black, Asian and older women were more often associated with unattractiveness by algorithms and stock photos, while photos of young white women were suggested more often as examples of beauty.[48] The researchers assume that the reason for the identified stereotypes could stem from a combination of the stock of available photos and characteristics of the search engines' indexing and ranking algorithms. The stock of online photos may reflect prejudices and biases of the real world, which are transferred from the physical world to the online world of the search engines.[49]

[28] It is known that engineers behind AI systems are in general white men. For instance, in the UK, less than 10% of computer programmers are women.[50] Or only 31% of all Google's US workforce are women.[51] It is known that the accuracy of AI systems is largely based on the training that an algorithm has undergone. If the training data is characterised, for instance, by the narrow view of a white man, the lack of diversity in this sector could lead to sexist and racist biases. In other words, the biases are incorporated – consciously or unconsciously – into algorithms and codes.[52] In the end, the existence of stereotypes in the online world can foster discrimination against minorities in both the online and the real world. A well-known example of the problem appears in facial recognition software: Conventional facial recognition works accurately if a person has a white skin colour. In this case, the software can identify both the skin colour and the gender in 99% of the cases. However, the darker the skin, the more difficult it is for the software to classify the gender.[53] The error rate is highest for dark-skinned women.[54]

5. AI regulation in Europe

5.1. «Artificial Intelligence Act» of the European Commission

[29] On 21 April 2021, the European Commission published a proposal for an «Artificial Intelligence Act» – a draft bill on the regulation of AI – in order to develop human-centric AI and eliminate mistakes and biases to ensure the AI is safe and trustworthy.[55] The proposal follows

48 Camila Souza Araújo/Meira Wagner/Virgilio Almeida, Identifying Stereotypes in the Online Perception of Physical Attractiveness, 2016, *passim* (https://arxiv.org/pdf/1608.02499.pdf).

49 *Ibid*; see also Jahna Otterbacher, New Evidence Shows Search Engines Reinforce Social Stereotypes, 20 October 2016 (https://hbr.org/2016/10/new-evidence-shows-search-engines-reinforce-social-stereotypes).

50 Aurore Lentz, Garbage in, garbage out: is AI discriminatory or simply a mirror of IRL inequalities?, 18 January 2021 (https://www.universal-rights.org/blog/garbage-in-garbage-out-is-ai-discriminatory-or-simply-a-mirror-of-irl-inequalities/; cit. Lentz).

51 Allison Levitsky, For the first time, white men weren't the largest group of U.S. hires at Google this year, 5 May 2020 (https://www.bizjournals.com/sanjose/news/2020/05/05/for-the-first-time-white-men-werent-the-largest.html).

52 Byrd Pinkerton, He's Brilliant, She's Lovely: Teaching Computers To Be Less Sexist, 12 August 2016 (https://www.npr.org/sections/alltechconsidered/2016/08/12/489507182/hes-brilliant-shes-lovely-teaching-computers-to-be-less-sexist?t=1617819757786); Goodman (Fn. 43); *Ibid*.

53 Karen Hao, This is how AI bias really happens – and why. It's so hard to fix, in: MIT Technology Review, February 2019 (https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/); Steve Lohr, Facial Recognition Is Accurate, if You're a White Guy, 9 February 2018 (https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html).

54 *Ibid*; Lentz (Fn. 50).

55 Saverio Puddu, EU takes a human centric approach to regulation of AI, 19 May 2021 (https://techinsights.linklaters.com/post/102gyih/eu-takes-a-human-centric-approach-to-regulation-of-ai).

the European Commission's «White Paper on Artificial Intelligence»[56] which was published on 19 February 2021 and represents the starting point for the regulation of AI in the EU. This historic step positions Europe as the first continent to uniformly regulate AI in addition to the existing data protection regulation. With the help of this regulation, the EU will set standards for the use of AI that may also have an impact beyond European borders.[57]

5.1.1. Objective of the draft bill on the regulation of AI

[30] The objective of the draft bill on the regulation of AI is to improve the functioning of the internal market by creating a single legal framework in particular for the development, marketing and use of AI in line with the values of the EU. The free cross-border movement of AI-based goods and services shall be ensured. Then, Member States will be prevented from restricting the development, marketing and use of AI systems in a way that is not stipulated in the regulation. The regulation pursues a high level of protection of general interests such as health, safety and fundamental rights. Citizens should be able to trust that the use of AI technology will be safe and in compliance with the EU laws. Furthermore, the regulation presents requirements to minimise the risk of discrimination through AI, in particular in relation to the quality of data sets used for the development of AI systems.[58]

[31] The draft law is similar in approach and structure in some ways to the GDPR, its regulatory approaches and concepts. It is in line with the EU's tendency to require more information about the algorithms used. The draft bill is then only one part of a planned EU regulatory package on AI. Thus, the topic of a revision of the General Product Safety Directive[59] and the Product Liability Directive[60], which is also suggested by the Commission in the White Paper and in the «Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics» of 19 February 2020[61], is not included in the current draft bill on AI regulation.[62]

---

56 White Paper on Artificial Intelligence of 19 February 2020, A European approach to excellence and trust, COM (2020) 65 (https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf).

57 Oliver Guggenbühl, Potenzial noch nicht ausgeschöpft – Ein Kommentar zur vorgeschlagenen KI-Regulierung der EU, 28 April 2021 (https://www.statworx.com/ch/blog/potenzial-noch-nicht-ausgeschoepft-ein-kommentar-zur-vorgeschlagenen-ki-regulierung-der-eu/; cit. Guggenbühl).

58 Saverio Puddu/Ana Isabel Rollán Galindo/Kay Firth-Butterfield, What the EU is doing to foster human-centric AI, 3 May 2021 (https://www.weforum.org/agenda/2021/05/ai-and-ethical-concerns-what-the-eu-is-doing-to-mitigate-the-risk-of-discrimination/).

59 Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on general product safety, OJ L 011, 15/01/2002 (https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32001L0095&from=EN).

60 Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, OJ L 210, 07/08/1985 (https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31985L0374&from=EN).

61 Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee of the 19 February 2020, Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics, COM(2020) 64 (https://ec.europa.eu/info/sites/default/files/report-safety-liability-artificial-intelligence-feb2020_en.pdf).

62 France Vehar/Jan Pohle, EU Kommission veröffentlicht Vorschlag des «Artificial Intelligence Act» – Die Kernpunkte des Game Changers zur Regulierung von Künstlicher Intelligenz im Überblick, 21 April 2021 (https://blogs.dlapiper.com/iptgermany/2021/04/21/eu-kommission-veroffentlicht-vorschlag-des-artificial-intelligence-act-die-kernpunkte-des-game-changers-zur-regulierung-von-kunstlicher-intelligenz-im-uberblick/#page=1; cit. Vehar/Pohle).

It is expected that the Commission will adopt a draft bill in the second quarter of 2021 that will address the revision of the General Product Safety Directive.

[32] The EU Commission has decided to classify AI as a product and regulate it via access to the market. The focus is on the providers of AI software and the users who use such software in their products or services. They are to be obliged to ensure that no harm is caused by AI. However, it is interesting that the current draft does not contain a direct basis for filing a complaint in the event of violated rights, as is the case in the GDPR.[63]

5.1.2. Material Scope

[33] The regulation is materially applicable to so-called «AI systems» (*see for definition chapter II. 2.1)*. For the purpose of the regulation, AI itself is defined as software that is developed with one or more of the techniques and approaches listed in *Annex I* and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with (Art. 3 (1)). Pursuant to Article 4 of the proposal the Commission is empowered to adopt delegated acts to amend the list in Annex I, in order to update that list to market and technological developments on the basis of characteristics that are similar to the techniques and approaches listed therein.

5.1.3. Categories of AI systems

[34] The European Commission has designed a pyramidal scheme that presents three categories of AI systems – split according to their potential risk. In particular, the regulation follows a risk-based approach, differentiating between uses of AI that create (1) an *unacceptable risk*, (2) a *high risk*, and (3) *low* or *minimal risk*. The majority of the regulation focuses on high-risk AI systems.

[35] The use of unacceptable, prohibited AI systems is considered unacceptable as contravening Union values, for instance by violating fundamental rights.[64] In particular, AI systems that are likely to cause physical or psychological harm to someone by manipulating their behavior, opinions or decisions are to be prohibited. In the draft, Art. 5(1) lists which AI systems are to be banned:

- «*AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm*» *(lit. a);*
- «*AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm*» (lit. b);

63 Ruth Fulterer, Die EU will künstliche Intelligenz regulieren: die wichtigsten Punkte und die Bedeutung für die Schweiz, in: NZZ, 10 May 2021 (https://www.nzz.ch/pro-global/technologie/die-eu-reguliert-ki-was-ist-wichtig-was-betrifft-die-schweiz-ld.1623518?reduced=true; cit. Fulterer).

64 Para. 5.2.2 of the Proposal AI Act (Fn. 13).

- Social-Scoring by public authorities or on their behalf, such as those launched in China,[65] *that evaluate or classify «the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics*» (lit. c);
- AI systems for the surveillance of individuals – with other words, «*real-time*» *remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement*. Exceptionally, however, such measures may be allowed under certain circumstances: e.g. «*the targeted search for specific potential victims of crime, including missing children*» (lit. d).

[36] Use cases that do not fall into the category of unacceptable, prohibited AI systems are to be placed on a risk scale.[66] Thus, depending on the risks associated with the AI system, the other three categories are subject to different regulatory requirements.

[37] Article 6 of the draft classifies AI systems as *high-risk* when (1) «*the AI system is intended to be used as a safety component of a product, or is itself a product, covered by the Union harmonisation legislation listed in Annex II*», and when (2) «*the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or putting into service of that product pursuant to the Union harmonisation legislation listed in Annex II*».[67] Annex II is a list of product safety and related market surveillance legislation. In addition, AI systems referred to in Annex III shall also be considered high-risk.[68] The Commission expects high-risk systems to be found in a variety of fields. Examples include AI systems for autonomous driving and drones, facial recognition, surgical robots and applications to sort CVs from job candidates.[69] For high-risk AI systems, there are considerable obligations for the operator of the AI system. These include the quality of the data and the accuracy of the AI system, as well as obligations for technical documentation, record-keeping, transparency, information to users and ensuring human oversight.[70] Before being placed on the EU market or being put into service, these AI systems must also undergo a conformity assessment to demonstrate that the systems are compliant with the requirements set out in the regulation.[71] In addition, the system shall be registered in the EU database.[72]

[38] For high-risk AI systems, the draft bill requires that training, validation and testing data sets shall be relevant, representative, free of errors and complete. These data shall be subject to appropriate data governance and management practices. Those practices concern, for example, the examination in view of possible biases.[73] To conclude, the draft is also intended to address the particular problem of inequalities and power asymmetries through AI systems described in this publication.

65 Wikipedia, Social Credit System (https://en.wikipedia.org/wiki/Social_Credit_System).

66 Vehar/Pohle (Fn. 62).

67 Art. 6(1) of the Proposal AI Act (Fn. 13).

68 Art. 6(2) of the Proposal AI Act (Fn. 13).

69 Jorge Liboreiro, «The higher the risk, the stricter the rule»: Brussels' new draft rules on artificial intelligence, 21 April 2021 (https://www.euronews.com/2021/04/21/the-higher-the-risk-the-stricter-the-rule-brussels-new-draft-rules-on-artificial-intellige).

70 Art. 8–15 of the Proposal AI Act (Fn. 13).

71 Art. 43 of the Proposal AI Act (Fn. 13).

72 Art. 51 of the Proposal AI Act (Fn. 13).

73 Art. 10 of the Proposal AI Act; Björn Finke, Was das neue KI-Gesetz der EU vorsieht, 21 April 2021 (https://www.sueddeutsche.de/wirtschaft/ki-ai-eu-gesichtserkennung-intelligenz-1.5271653).

[39] For AI systems with *low* or *minimal risk*, specific transparency obligations will apply in order to allow users to make informed decisions and to be aware that they interact with an AI system, for example with a chatbot.[74]

5.1.4. Criticism of the proposal for AI rules

[40] The idea of the EU to promote the development of AI and at the same time to control it in the right way is to be welcomed from the author's point of view. AI applications should not lead to the restriction or even elimination of fundamental rights and freedoms that are essential for a constitutional state. With the proposal, the EU recognises that AI is groundbreaking for the future of the European market. Guidelines for a technology with such great implications are important – as is the promotion of innovation.[75]

[41] However, the draft will likely undergo significant changes in the course of the EU legislative process before it enters into force as law. In particular, the allocation of AI systems to the respective risk classes could be improved. For example, the proposal does not contain a clear definition of «high risks». Since developers themselves are responsible for assessing their applications, a clearly defined scale for assessing risks would be essential. Articles 6 and 7 do describe risks and give examples of «high risk applications», but a process for assessing risks of an AI application is not defined.[76] The current draft could then also be modified, for example, to include mandatory auditing by an independent auditing organisation for all high-risk applications and not only for certain AI systems with high risks.[77]

[42] Furthermore, according to the draft, the training, validation and testing of datasets must be subject to appropriate data governance and management practices, also with a view to possible bias. Providers of high-risk continuous learning AI systems must therefore ensure that potentially biased outputs are equipped with appropriate mitigation measures when they are used as inputs in future operations. Here, the AI regulations are unclear about how AI systems will be tested for potential bias, in particular whether the benchmark will be equality of opportunity or equality of outcome.[78]

[43] Finally, in the case of *unacceptable* AI, it is noticeable that in each case a relatively large number of cumulative conditions must be met for the prohibition norm to actually intervene. For example, not all AI applications that manipulate children's behavior through subliminal techniques are prohibited. Rather, the prohibition only applies if the danger of physical or psychological harm is caused.

74 Art. 52 of the Proposal AI Act (Fn. 13).

75 Vehar/Pohle (Fn. 62).

76 Guggenbühl (Fn. 57).

77 Algorithm Watch, AlgorithmWatch's response to the European Commission's proposed regulation on Artificial Intelligence – a major step with major gaps, April 2021, p. 3 (https://algorithmwatch.org/en/wp-content/uploads/2021/04/AWs-response-on-ECs-AI-regulation-proposal-April-2021-v1-2021-04-21.pdf).

78 Julia Wilson/Roderick Beudeker/Autumn Sharp, New Draft Rules on the Use of Artificial Intelligence, 14 May 2021 (https://www.bakermckenzie.com/en/insight/publications/2021/05/new-draft-rules-on-the-use-of-ai).

5.2. Impact on Switzerland

[44] The EU regulation on AI extends far beyond the borders of the EU. The regulation applies to all products that are placed on the market in the EU or that affect people in the EU. Especially in the software sector, where new products are costly to develop but very cheap to reproduce, such rules can quickly have an impact in other countries, including Switzerland.[79] Most AI providers will not develop their own products for Switzerland, hence new European standards will have an impact in this country as well. This is exactly what happened with the data protection regulation (GDPR).

[45] In Switzerland, the Federal Council made AI a core theme of the so-called *Digital Switzerland Strategy* in 2018.[80] Therefore, an interdepartmental working group on AI was set up. In December 2019 the group published a report in which it explained the challenges regarding AI for Switzerland. The report states that relevant legal principles in Switzerland would usually be formulated in a technology-neutral way so that they could also be applied to AI systems. In concrete, it is stated that the existing legal framework already permits and limits the use of AI in principle (e.g. Federal Act on Gender Equality), and also applies in particular to discrimination that may arise as a result of AI decisions. Thus, there would be no need for fundamental adjustments to the legal framework. However, given technological dynamics, it could not be ruled out that this assessment could change quickly. In 2020, the same interdepartmental working group then developed guidelines on the use of AI within the Federal Administration, meaning a general frame of reference for federal agencies and external partners entrusted with governmental tasks. The guidelines were adopted by the Federal Council in November 2020.[81]

[46] In a response to a parliamentary motion by National Councillor Bellaîche from 10 June 2021, in which Switzerland's participation in the European regulation of digitalization was discussed, the Federal Council pointed out that it was closely observing the relevant developments in the EU and their impact on Switzerland.[82] The Federal Council had been informed of new results of an analysis on 30 June 2021 – the analysis was published on the OFCOM website on 8 July 2021 and generally revealed that the EU digital strategy measures already initiated by the EU will also be relevant for Switzerland in the future. Among other things, the analysis states – with regard to the topic of AI regulation – that the general legal framework in Switzerland is currently considered to be sufficient to deal with the challenges that AI poses. Nevertheless, the Federal Council would be aware that the new draft of the EU AI Act would have an extraterritorial effect and would therefore also have an impact on Switzerland and the companies and research institutions operating in this country. For this reason, the Swiss government would closely follow

79 Fulterer (Fn. 63).

80 The Federal Council, Press Release, New guidelines for digital Switzerland, 6 September 2018 (https://www.bakom.admin.ch/bakom/en/homepage/ofcom/ofcom-s-information/press-releases-nsb.msg-id-72053.html).

81 Allesandro Curioni et al., Recommendations for an AI Strategy in Switzerland A white paper organised by the SATW topical platform on Artificial Intelligence, December 2019, p. 4 et seqq. (https://www.satw.ch/fileadmin/user_upload/documents/02_Themen/08_Kuenstliche-Intelligenz/SATW-Swiss_AI_Strategy.pdf); The Federal Council, Guidelines on Artificial Intelligence for the Confederation General frame of reference on the use of artificial intelligence within the Federal Administration, 25 November 2020 (https://www.sbfi.admin.ch/sbfi/en/home/eri-policy/eri-21-24/cross-cutting-themes/digitalisation-eri/artificial-intelligence.html).

82 Motion 21.3676 (https://www.parlament.ch/de/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20213676).

the developments of the legislative process within the EU in order to be able to take measures at an early stage if necessary.[83]

[47] Although Swiss policymakers are becoming increasingly aware of the importance of AI, the debate on AI in Switzerland so far has been rather poor. Yet Switzerland should not fear the challenge of making the use of ADM systems and the algorithms themselves fairer. Indeed, the problem of algorithmic bias may even represent an opportunity for Switzerland – it could position itself as a role model internationally.[84] In my opinion, the Swiss policymakers will be under increasing pressure to discuss the subject of AI more intensively and in greater depth in public – especially because Switzerland maintains a close exchange with the EU, which has now definitely launched the debate with its «AI Act».

[48] The aim of a legislative regulation of AI in Switzerland could be to minimise the risks associated with AI applications in certain sectors (e.g. use of AI in the health sector), to ensure sufficient transparency and to provide the necessary tools to enable affected persons to take action against specific disadvantages. In this case, it would probably make sense to make selective adjustments to the existing regulations in the affected legal fields. In addition, adjustments to generally applicable norms or a broader interpretation of these legal norms will probably be necessary in the future. For example, the academic community proposes the implementation of a general prohibition of discrimination, in the event that discrimination by private persons (and in particular by their AI) is not sufficiently covered by the corresponding interpretation of the protection of legal personality under civil law (Art. 28 of the Swiss Civil Code).[85]

5.3. WHO guidance on ethics and governance of AI for health

[49] Finally, the World Health Organization (WHO) has also started to focus on AI and published its guidelines on the use of AI in the context of health on 28 June 2021. Its guidance on *Ethics & Governance of Artificial Intelligence for Health* is the product of eighteen months of deliberation amongst leading experts in ethics, digital technology, law, human rights, as well as experts from Ministries of Health. According to the guidance, new technologies that use AI must put ethics and human rights at the heart of its design, deployment, and use. The report identifies the ethical challenges and risks with the use of AI of health and presents six consensus principles to ensure AI works to the public benefit of all countries. The recommendations shall ensure that the governance of AI for health maximises the promise of the technology and, on the other side, holds all stakeholders accountable and responsive.[86]

[50] The report endorses a set of key ethical principles for the use of AI for health which shall be used as a basis for governments, technology developers, companies, civil society and intergovernmental organisations – in order to limit the risks and maximise the opportunities intrinsic

83 Digital Switzerland (https://www.bakom.admin.ch/bakom/en/homepage/digital-switzerland-and-internet/strategie-digitale-schweiz/digitale-schweiz.html).

84 Alexis Perakis/Maya Guido/Amir Mikail, Algorithmic Bias in der Schweiz, 21 April 2021 (https://reatch.ch/publikationen/algorithmic-bias-in-der-schweiz).

85 Nadja Braun Binder/Thomas Burri/Melinda Florina Lohmann/Monika Simmler/Florent Thouvenin/Kerstin Noëlle Vokinger, Künstliche Intelligenz: Handlungsbedarf im Schweizer Recht, in: Jusletter 28 June 2021, para. 54–55.

86 Ethics and governance of artificial intelligence for health: WHO guidance, 28 June 2021 (https://www.who.int/publications/i/item/9789240029200; cit. WHO guidance of AI for health).

to the use of AI for health. Two of the six principles are of particular interest in the present discussion:

- **Ensuring transparency, explainability and intelligibility**
  «*AI technologies should be intelligible or understandable to developers, medical professionals, patients, users and regulators. Two broad approaches to intelligibility are to improve the transparency of AI technology and to make AI technology explainable. Transparency requires that sufficient information be published or documented before the design or deployment of an AI technology and that such information facilitate meaningful public consultation and debate on how the technology is designed and how it should or should not be used. AI technologies should be explainable according to the capacity of those to whom they are explained.*»[87]
- **Ensuring inclusiveness and equity**
  «*AI technologies should not encode biases to the disadvantage of identifiable groups, especially groups that are already marginalised. Bias is a threat to inclusiveness and equity, as it can result in a departure, often arbitrary, from equal treatment. (. . .) AI tools and systems should be monitored and evaluated to identify disproportionate effects on specific groups of people. No technology, AI or otherwise, should sustain or worsen existing forms of bias and discrimination.*»[88]

[51] The report provides practical advice for implementing the WHO guidance for three different stakeholders: AI technology developers, ministries of health, and healthcare providers. At this current stage, the considerations in the guidance are intended only as a starting-point for context-specific discussions and decisions by the stakeholders. In the coming months, the WHO will focus on developing an implementation plan. However, collective action will ultimately be required for the implementation of the WHO guidance as well as its success.

6. Quo vadis?

[52] AI systems are becoming increasingly important in the health sector and offer a wide range of possibilities that indeed have the potential to revolutionalise healthcare. However, AI systems can reflect bias and discrimination in several ways: firstly, in patterns of pre-existing health discrimination that then become entrenched in datasets; secondly, in the representativeness of the data; and finally, in human decisions made during the design, development and use of these AI systems. Furthermore, there is the black box problem. The process behind the output of a DL model is not transparent and can therefore not be reconstructed. Even the coders themselves, who developed the system, have difficulties to understand how the output was generated in a DL system.

[53] AI discrimination in healthcare is a serious problem that can harm patients. For instance, if a training data set only contains data with limited diversity (e.g. WEIRD data), the system will not be able to build a universal set of rules and people who do not correspond to the «standard human» will be discriminated.

87 WHO guidance of AI for health (Fn. 86), pp. XIII and 26–27.

88 WHO guidance of AI for health (Fn. 86), pp. XIII–XIV and 29–30.

[54] In order to ensure that in the future AI will not produce more biased output and reinforce the inequalities that already exist in Health Tech, several options need to be addressed: It should be ensured that clinical research is conducted on diverse study populations. Then, more people from minorities or, for example, more women should be recruited to develop AI tools. Furthermore, the outputs of AI should be regularly reviewed and cross-checked with the diagnoses of humans. Finally, healthcare providers should keep the risk of bias in mind when selecting ADM systems. In this regard, there must be more transparency in the design process of AI systems, e.g. with periodical reviews and assessments, in order to adequately monitor whether AI is safe (non-maleficent) and effective (beneficent).[89]

[55] As AI systems continue to advance not only in medicine but also in other areas of everyday life, people should be able to trust them – trustworthiness is a prerequisite for their acceptance. This opinion is shared by Margrethe Vestager, Executive Vice-President for a Europe fit for the Digital Age, particularly with regard to the draft bill on the regulation of AI: «*On Artificial Intelligence, trust is a must, not a nice to have. With these landmark rules, the EU is spearheading the development of new global norms to make sure AI can be trusted. By setting the standards, we can pave the way to ethical technology worldwide and ensure that the EU remains competitive along the way. Future-proof and innovation-friendly, our rules will intervene where strictly needed: when the safety and fundamental rights of EU citizens are at stake.*» Because the EU attaches great importance to certain values and the rule of law, efforts will be made to further make AI systems suitable for a safe, reliable and unbiased use with clear rules.

[56] AI and its related problems have to move even more into the focus of science, politics and the media. An open discussion at the interface of the aforementioned areas is becoming indispensable in this context. This will allow existing problems to be identified and the key next steps for controlling them to be discussed and tackled. Algorithmic biases could, for example, be minimised through ongoing research and data collection which is representative of a diverse population. After all, if AI is indeed fair, it can help us overcome and rethink our own societal biases. AI might reveal existing discriminatory practices and trigger reflection on decision criteria that underlie them. Thus, in the author's view, it is worthwhile to progress new ethical frameworks for AI and to regulate AI as efficiently and purposefully as possible, in order to ultimately prevent inequalities and power asymmetries, e.g. in the context of health tech, and even shed light on the existing biases we have as humans.

Anne-Sophie Morand, Dr. iur., attorney-at-law.

The author thanks Amir Mikail for the valuable feedback on the technical aspects of this publication. He studied mechanical engineering at ETH Zurich, also holds a Master's degree in Science, Technology and Policy from ETH and works as a consultant.

89 Hoffman (Fn. 31); Lentz (Fn. 50).