IEAI
Technical University of Munich
Munich Center for Technology in Society
Institute for Ethics in Artificial Intelligence

Research Brief – October 2021

Assessing Fairness in AI-enabled Public Health Surveillance

By Ellen Hohma

With the spread of the Covid-19 pandemic and the intensified need to contain the virus, new technologies exploring untraditional data sources and equipped with innovative AI techniques have been proposed and quickly developed, often without a significant public debate on the ethics of their use. Particularly, aspects of fairness are often not sufficiently considered, as building sound and just algorithms usually entails weighing numerous and subjective parameters. This omission has the potential to cause significant risks to the effectiveness of AI-enabled health surveillance tools and with that the well-being of the broader society. This brief explores the concept of fairness as it relates to the example of health surveillance systems, as well as how the concept interacts with other ethical considerations applied to AI-based tools.

The outbreak of the Covid-19 pandemic has proven to many in society the importance of a continuous and conscientious tracking of health threats for the safety and well-being of the public. A useful means to identify and confront such crises are health surveillance systems. The World Health Organization (WHO) defines public health surveillance as "the continuous, systematic collection, analysis and interpretation of health-related data" (World Health Organization, 2021c). With the ongoing creation and amplification of massive amounts of data, the development of specialized data analysis tools for the health surveillance sector has likewise increased. Innovative technologies engaging Machine Learning (ML) and Artificial Intelligence (AI) have been proposed to monitor and improve health standards.

With the opportunity to collect data on a large scale and the creation of Big Data processing techniques, however, conflicts arise concerning health surveillance norms. Concerns around how such data should be evaluated and to what extent interventions can be planned and enforced upon it have quickly spread throughout society. Especially fairness issues have raised major discussions. The need for self-isolation to contain virus spread has increased considerations on the justification, fairness and viability of such interventions. Currently, the debate concerning which rights to grant vaccinated people, while avoiding discrimination of non-vaccinated individuals, has become very prominent. With the development of new technologies, the need to agree on a common set of principles has intensified, as designing such tools requires a translation from publicly acceptable norms to tangible and implementable rationales. More specifically, given that fairness is a subjective feeling, its implementation into technologies and AI systems is complex. A clear and careful investigation, as well as concrete guidance is needed to ensure fairness of AI-enabled tools, especially in the health surveillance segment, as its impacts on public society are manifold.

In this Brief, we will first outline how the field of public health surveillance has evolved from mere incident recording to complex AI-based prediction systems. We will further discuss possible fairness issues and their roots from a technical point of view. Finally, we will highlight the potential for interplay but also conflicts between fairness and other ethics principles in public health surveillance technologies.[1]

Opportunities of AI for health surveillance technology

Epidemiologic monitoring and intervention have a long tradition throughout human history, with the first documented epidemic recording dating back to 3180 B.C. (Choi, 2012). While first health surveillance actions focused on the observation of diseases and reporting of resulting deaths, the need for analyzing the obtained data and its potential for disease control soon became apparent (Teutsch & Churchill, 2000). A prominent contributor to advancing public health surveillance, especially linking investigations to interventions, is the anesthesiologist John Snow (Choi, 2012). Studying the course of one of four worldwide cholera outbreaks between 1817 and 1875, Snow hypothesized the root of a London disease outbreak in 1849 to be fecal-contaminated water supplies (Choi, 2012; Gerstman, 2013). By mapping cholera death cases and evaluating their water gathering habits, he identified the outbreak's source to be a public water pump on Broad Street. Removing the pump handle, and hence preventing the public from collecting water from this well, supported his theory as the local disease outbreak waned (Choi, 2012).

With modern technological achievements, the opportunities and potential for health surveillance and disease control naturally advanced. Epidemic and health monitoring developed from typically locally concentrated, individual observations and recordings to large scale, structured, preventive data collection and analyses (Lee & Thacker, 2011). Today, most countries operate national health agencies in order to monitor diseases and

[1] The research in this Brief is based upon the IEAI Project "Building Strategic Partnerships to Understand Ethics and the Use of AI to Manage Health Related Crises" with Principal Investigators Prof. Christoph Luetge, Dr. Caitlin Corrigan and Prof. Dr. Rafael A. Calvo.

population health, with the World Health Organization (WHO) centralizing and assembling domestic activities on a global level, as well as directing and coordinating international health work (World Health Organization, 2021a). Information feeding the investigations is drawn from a variety of data sources, including regularly repeated health surveys or disease registries, for example the monitoring of cancer occurrences (Lee & Thacker, 2011). Traditionally relevant information extracted merely from hospital and laboratory report shifted to incorporate syndromic data, such as the number of patients visiting the emergency department (Chiolero & Buckeridge, 2020).

With the continuous acceleration of data production and recent advancements in Big Data processing, further unconventional data sources became the subject of investigations. For instance, in an effort to improve detection and tracking of influenza outbreaks, Google launched Google Flu Trends in 2008, a publicly available web service which communicated predicted influenza occurrences based on a linear regression model of multiple, distinct Google search entries, highly correlating with the influenza time series (Aiello et al., 2020). Supported by the US national public health agency, the Center for Disease Control and Prevention (CDC) developed an algorithm which was able to predict influenza-like disease outbreaks one to two weeks ahead of traditional systems (Aiello et al., 2020; Lee & Thacker, 2011). Besides search engine query data, other non-health-related data sources, such as social media (particularly Twitter geolocation data), were found to be a valuable means for identifying potential illness risks. Rocklöv et al. (2019) proposed to use Twitter geodata along with flight passenger data to trace and predict the spread of the Chikungunya virus in the Mediterranean area.

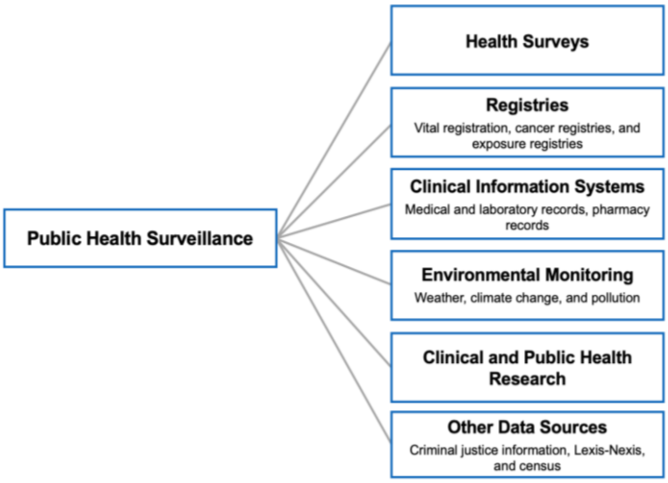
Public Health Surveillance
Health Surveys
Registries
Vital registration, cancer registries, and exposure registries
Clinical Information Systems
Medical and laboratory records, pharmacy records
Environmental Monitoring
Weather, climate change, and pollution
Clinical and Public Health Research
Other Data Sources
Criminal justice information, Lexis-Nexis, and census

***Figure 1**: Information Sources for public health surveillance. Source: own representation based on Lee & Thacker (2011)*

Through the inclusion of new input sources and increased gathering of massive amounts of data the need for innovative techniques enabling large-scale analysis becomes evident. In particular, the interdisciplinary and collaborative work of multiple stakeholders from hospitals, pharmaceutical industries, local and national authorities which is required for effective public health surveillance has resulted in an increased need to apply novel approaches of automation (Neill, 2012). Several purposes and potential for AI-supported health surveillance tools have been proposed and studied. Text mining, for instance, can serve as a powerful instrument to integrate diverse data types and enable an automated and more cost-efficient evaluation and extraction of disorganized information from electronic health records (Bi et al., 2019). In this manner, Murff et al. (2011) applied Natural Language Processing (NLP) to automatically determine postoperative complications from textual clinical documents.

Natural Language Processing (NLP) involves teaching computers how to interpret human language, as well as retrieve and understand inherent content and contexts.

Tracking and predicting disease spread is especially relevant when applying ML methods in health surveillance. One such example is the usage of random forests to identify disease transmission by mapping data from geospatial applications and retrieve estimations of infection probabilities (Bi et al., 2019). Similarly, Support Vector Machines (SVMs) can be used in the prediction and forecasting of epidemics. Thomson et al. (2006) use such ensemble methods to create a malaria pre-warning system based on seasonal climate forecasts. While their practical application in public health surveillance currently still lacks behind other fields (Rocklöv et al., 2019), new technologies, data sources and methodologies evidently bear promising potential for epidemiologic monitoring and intervention.

Fairness and its absence as a risk for AI-based health surveillance

The shift towards individually collected data, with the help of recent technologies, opens room for new epidemiological tracking and disease control. Precision of forecasted health-related events increases with the collection of more detailed information. Traditionally, people who did not or could not afford to visit a doctor were not captured by classic medical or laboratory reports. Using new data sources, like social media or Google queries, can largely increase the representation of such groups (Mello & Wang, 2020). While using AI to analyze new data sources can improve health surveillance systems as such, simultaneously, disparities in characteristics and demands of certain groups can be recognized and thus chances for need-based support can be balanced. However, conversely, it illustrates the necessity for fairness in AI-based public health surveillance, as a lack of equal representation and opportunities can lead to malfunctioning and hence unpredicted health risks.

One of the risks for effective performance of health monitoring systems is the inadequate data collection strategy, resulting in an unsatisfactory representation of particular groups (Gasser et al., 2020; Klingler et al., 2017; TUM Institute for Ethics in AI, 2020). This risk intensifies if data is taken from initially non-health-related sources, requiring internet and technical devices (Budd et al., 2020; Klingler et al., 2017; Mello & Wang, 2020). Around 33% of the world's population does not have a mobile service subscription (GSMA, 2021). Additionally, considering the case of social media, the mere access to the internet does not mean all population groups equally use them. The prevailing majority, 82.9%, of worldwide twitter users, for instance, is below the age of 50 (Statista, 2021). Especially for healthcare purposes, a considerable target group is hence underrepresented in this dataset. While new population groups can be reached, it must be ensured that others are not excluded in return to enable effective disease prevention and control.

However, the opposite extremum of poor data collection similarly incorporates risks. Discrimination and stigmatizing of certain groups can be the result of over-representation and an unjustified focus on specific target populations (Gasser et al., 2020; Klingler et al., 2017). New technologies collecting and analyzing large amounts of data can be used to gather personal information, such as ethnicity or race. There is a risk that such data will influence health-related decisions. Linked with knowledge about a person's health status, stigmas might evolve as observed during the beginning of the COVID-19 pandemic, where an increase of discrimination against south-east Asians could be perceived (Gasser et al., 2020).

On a global scale, further risks arise with a potentially biased distribution of innovative health surveillance technology between regions or countries. Inadequate priority setting can pose a serious risk to overall health surveillance performance. Spatial decisions may impact the tools' efficacy if areas that are more relevant to richer countries are favored over areas of high need (Klingler et al., 2017). In 2019, the Global Health Security (GHS) Index published a comprehensive assessment of health security preparedness of 195 countries and revealed severe deviations in regards to a country's income (Cameron et al., 2019).

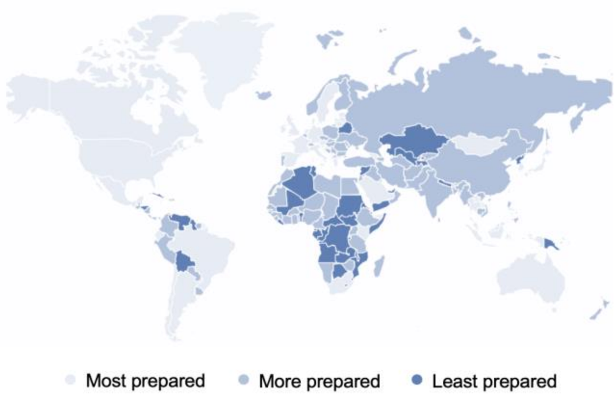

***Figure 2**: Preparedness level on 'early detection and reporting epidemics of potential international concern' per country.*
Source: own representation based on Cameron et al. (2019)

Besides other features, a country's ability for "early detection and reporting [of] epidemics of potential international concern" (Cameron et al., 2019) was studied. The 34 countries characterized as low income reached an average score of 30.87[2], with Zimbabwe (65.5) scoring highest. For the 60 high income countries, the average score of 55.56 was

[2] On a scale ranging between 0 and 100, with 100 indicating perfect preparedness regarding disease detection and reporting mechanisms

substantially higher, with the United States reaching the highest score of 98.2. The risk of disproportionate focus on wealthier regions in epidemiologic monitoring can lead to missing relevant health disturbances thus affecting public health. Further, a waste of resources due to wrong emphasis in the development of surveillance systems needs to be prohibited (Klingler et al., 2017). Individual interests of groups need to take second place to reaching an equal distribution of benefits and burdens between the disproportionally developed nations. This should mitigate involved risks and ultimately serve the goal of global health.

The root of unfairness in AI systems and rising countermeasures

With their many risks involved, discriminative surveillance systems, and the need to avoid them, leads to the question of how prejudices arise within AI technology. The most commonly known problem is bias incorporated in the data used to train an AI system. Any AI-based technology requires data to draw conclusions, come to a decision or calculate predictions. However, the underlying data is rarely neutral as it is produced by humans who include objective perceptions and opinions (Chouldechova & Roth, 2018). An extensive summary of potential sources and types of biases is provided by Mehrabi et al. (2019). While some data biases are created willingly, many are unconsciously introduced. Population bias, for instance, incorporates disparities between demographics or representation of certain groups in the dataset and the original target population (Olteanu et al., 2019). Selected training data must fit to the context and target population in terms of, for example, culture, socio-demographic background, as well as behavior (Feuerriegel et al., 2020). While such deviations might be detectable and removable, other influences such as observer bias, where preferences are included, because researchers unintentionally project their assumptions into the data, are less obvious (Mester, 2017).

Besides biased datasets, further fairness issues can emerge from the algorithmic design. The underlying model chosen for an AI system is a crucial factor for determining fairness of outcomes (Feuerriegel et al., 2020). Obermeyer et al. (2019) hypothesized that the choice of surrogates selected for real-world concepts can significantly contribute to introducing new biases. They found that black people were discriminated against, receiving additional health insurance payments in a widely used algorithm because historic healthcare costs were chosen as the one of the determinants for the degree of illness.

To create a comprehensive framework targeting potential sources of algorithmic unfairness, researchers within the domain of Machine Learning have started to gather and summarize different definitions of fairness. Mehrabi et al. (2019), for instance, concluded 10 definitions that can help developers to implement impartiality into AI algorithms. These models range from ensuring equal opportunities for all groups, to predicting similar outcomes for similar individuals. However, selecting the right fairness model fitting a certain context remains a challenge. With the many nuances and individual, potentially mutually excluding perceptions fairness incorporates, the choice of technically implementable fairness approaches is not always obvious.

Data scientists and developers have started to build innovative toolkits and systems identifying biases in datasets and mitigating negative impacts from unfair algorithms.

As the issue of fair AI systems is not limited to the field of public health surveillance, much research has already targeted the problem of unfairness in AI. Data scientists and developers have started to build innovative toolkits and systems identifying biases in datasets and mitigating negative impacts from unfair algorithms. One such example is the IBM AI Fairness 360 (AIF360) toolkit (Bellamy et al., 2018). It aims at facilitating the integration of concepts from research in the field of AI fairness into industry standards to create a unified framework for fairness evaluation of algorithms. More concretely, AIF360 consolidates bias detection metrics, bias mitigation algorithms, as well as bias explanation measures to inform the users about potential impacts of identified objectives (Bellamy et al., 2018). A similar approach is endorsed by Fairlearn, proposing an open-source toolset that combines unfairness reduction methods with an interactive visualization dashboard (Bird et al., 2020). While these toolboxes provide effective and convenient instruments, as well as libraries to programmatically target unfairness issues, further sources, especially social objectives, such as differing fairness perceptions and values, cannot be merely targeted using technology.

In response, social scientists have reaffirmed the increasing obligation for extended research and several organizations and researchers have analyzed a variety of ethical challenges in public health surveillance, before and especially with the spread of the COVID-19 virus (e.g., Gerke et al., 2020; Lee, 2019; Lee et al., 2010; Lucivero et al., 2020; Nuffield Council on Bioethics, 2020; World Health Organization, 2020). These findings resulted in multiple, largely overlapping ethical guidelines and recommendations for the use of AI in public health surveillance (e.g., CNECT, 2020; Fairchild et al., 2017; Morley et al., 2020; World Health Organization, 2017), mainly along the 5 ethical principles of AI in society as proposed by Floridi et al. (2018): beneficence, non-maleficence, autonomy, justice and explicability.

Besides protecting privacy and data security, this leads to questions like how to ensure the proportionality of the introduced technology as opposed to the people's privacy protection or whether to maintain people's individual freedom to use such tools (TUM Institute for Ethics in AI, 2020). Although none of these guidelines focused solely on fairness issues, ensuring equal and just treatment is always fundamental. Funneling global efforts, the World Health Organization (WHO), for example, published a report that proposes 17 guidelines summarizing ethical issues in public health surveillance (World Health Organization, 2017). A particular focus regarding fairness lies on the global collaboration and mutual support in case of imbalanced resource access. Especially with the spread of the coronavirus and the extended need for innovative appliances to oppose the pandemic, such as contact tracing apps, AI ethics research aims are increasingly directed towards their investigation and regulation. Morley et al. (2020), for instance, suggest a checklist for the development of contact tracing apps. Fairness is emphasized in terms of equitable availability and equal accessibility of the developed tools. While in theory ensuring fairness is a commonly acknowledged goal, its practical application in health surveillance technology is not always obvious. Especially if fairness should be technically embedded in the application's design, a clear and tangible definition is crucial. However, agreeing on a prevailing and publicly accepted fairness definition that is manageable enough to be implementable, is still challenging. Often context-specific decisions that differ per situation are required to reflect individual, personal perceptions and also allow AI enabled tools simultaneously to serve public values. Therefore, a variety of research efforts is still needed that is precisely dedicated to comparing AI surveillance use in specific contexts in order to understand fairness perceptions in disparate situations.

The principle of beneficence entails preserving and facilitating public good, human dignity and a sustainable environment (Floridi et al., 2018).

Interplay and conflicts between fairness and other ethics principles

Fairness is one major fundamental in ethical frameworks for evaluating the impact of AI technologies, however, not the sole one. Multiple schemas have been developed to structure ethical issues and guide investigations concerning potential social conflicts. While some of them were explicitly designed for studying ethics in healthcare (e.g., Beauchamp & Childress, 2001; Marckmann et al., 2015), a few focus even further on specifically public health surveillance, as well as digital and AI-based tools used in this domain (e.g., Aiello et al., 2020; Floridi et al., 2019; Gasser et al., 2020; World Health Organization, 2017, 2021b). Essentially, they all incorporate and reflect similar social values. Therefore, to increase comparability with other industries and research fields, we will abide by a more general framework by Floridi et al. (2018) in this analysis.

Floridi et al. (2018) introduce the idea of fairness in their principle of justice, defined as "promoting prosperity and preserving solidarity" (p. 698). Similar to the concept of fairness, this includes avoiding discrimination and bias, as well as furthering equal opportunities and shared benefits. Besides ensuring justice, Floridi et al. (2018) propose 4 main principles: beneficence, non-maleficence, autonomy and explicability. As all 5 principles equally serve the same common good, to preserve human values and protect human dignity, surely conflicts may arise if several principles should be maintained simultaneously. These conflicts need to be consciously investigated and carefully weighed to respect and increase overall public well-being.

Beneficence

Fairness and beneficence can complement each other as more fairness can lead to greater social good. For the case of public health surveillance, a fairer data collection, i.e., a greater emphasis on appropriate representativeness of collected data, increases data accuracy and, hence, prediction quality. If characteristics and needs of minorities are better represented in the dataset, algorithms can respect them in the analysis and draw more concrete conclusions. Ultimately, it increases overall effectiveness of health surveillance technologies, and serves public well-being. From a global perspective fairness can similarly promote health monitoring systems and thus health protection. A systematic and comprehensive data collection is important to capture all health-related events. Ensuring a globally balanced allocation of health surveillance resources and opportunities can increase the chances for identifying diseases. The Covid-19 pandemic has shown once more that local outbreaks can quickly turn into worldwide threats. Transnational collaboration and exchange of knowledge and tools strengthens the mechanisms for jointly fighting diseases.

> ***While beneficence ensures that AI is used to ‘do only good’, the principle of non-maleficence aims at limiting AI applications to ‘do no harm’ (Floridi et al., 2018)***

However, potential conflicts arise between beneficence and fairness of the distribution of resources as they are often scarce. With the spread of the Covid-19 pandemic, many companies and organizations agreed to openly exchange knowledge and technology to jointly contribute to the common goal of fighting the disease. Apple and Google, for example, teamed up in a joint effort to support governmental and health agencies with the utilization of recent Bluetooth technologies for contact tracing (Apple, 2020). Such free information transfer cannot be expected in non-crisis times, nurturing the discussions about what a fair resource distribution entails and how it can be harmonized with the aim of expanding positive impacts. Serving the goal of maximizing health standards around the globe to improve overall well-being, for instance, would require a non-restricted sharing of disease surveillance tools. While this can be considered as a fair resource allocation, acknowledging and respecting achievements as well as the intellectual property of people developing a technology outlines another notion of fairness. Thus, defining fairness and a just sharing of health surveillance resources further incorporates weighing various benefits and interests of different stakeholder groups.

Non-Maleficence

The two principles of non-maleficence and beneficence are distinct in that evaluating beneficence of AI applications investigates how to facilitate and promote their positive outcomes, while ensuring non-maleficence seeks to constrain negative impacts of such tools and, if required, to restrict them. Fairness and non-maleficence therefore have a common goal of prohibiting personal harm through avoiding unfair treatment and discrimination.

However, simultaneously maximizing both principles can lead to controversies and potential mutual constraints. Aiming at equally improving data representativeness, more data, especially on possibly sensitive characteristics, needs to be collected and analyzed with regard to avoiding discriminative influences and a hyper focus on those features. However, this would increase the amount of data collected, hence, potentially invade a user’s data privacy. An example for the conflict between information retrieval and data privacy is the German contact tracing tool Corona-Warn-App. It was designed with data privacy as one of the highest priorities to inhibit any inference of movement patterns through the application (Laaff, 2020; Robert Koch-Institut, 2021). Concealing where exposures might have occurred thus protects infected people from exclusion and reproach. At the same time, not collecting and revealing such information essentially impairs and complicates the contact tracing work of health authorities. In the case of an exposure signaled by the app, it is not possible to determine whether an actual risk of being infected exists or whether the encounter took place with further hygiene precautions, for example outdoor or behind a plastic shield (Laaff, 2020). Accordingly, the health authorities must decide to some extent randomly how to proceed in such cases. This impairs fairness as individuality and appropriateness of the imposed measures cannot be guaranteed. Allowing the collection of more data could help make more effective assumptions and predictions about the pandemic, but in return would affect data freedom and privacy. Since resolution of this conflict is not entirely obvious, the balance between the two principles must be carefully studied and defined.

Autonomy

Ensuring a suitable balance between power handed over to the machine and control remaining with humans is an essential objective for creating responsible and trustworthy AI systems. In its relation to fairness, autonomy could serve as a catalysator for improving overall satisfaction with health surveillance tools, as it was found to moderate the relationship between fairness and satisfaction (Haar & Spell, 2009). In other words, increasing both autonomy and fairness could facilitate user satisfaction with health surveillance tools.

While a simultaneous promotion of both autonomy and fairness could hence endorse user acceptance and uptake of suitable technologies, unrestricted freedom of choice cannot always be guaranteed in public health surveillance. To enable responsible and adequate health protection, certain information must be shared with responsible authorities to allow for appropriate interventions. The German Infection Protection Act (IfSG) specifies in § 6 that, for example, measles, chickenpox and recently also Covid-19 must be reported to the national health agency if identified by a doctor or laboratory. To protect the broader society, although seemingly personal, patients cannot decide to withhold such disclosure. With the introduction of new technologies, the question prevails to what extent automating reporting and identification of diseases is acceptable, desirable or contributes to a responsible society. The inclusion of non-traditional data sources initially not used for retrieving health information, such as social media data from Twitter, can increase the representation of certain population groups and, hence, improve the comprehensive tracking of reportable diseases. Conversely, it might stifle user autonomy as a Twitter user might not be aware and, especially, has not consented to the use of their data for health surveillance purposes. With regard to data analysis, the question of how much power can be handed to artificial systems is even more pressing. While a reduction of human interference could limit the influence of subjective human biases, the extent to which AI-based health monitoring systems can suggest or even enforce interventions needs to be clarified. Especially given that theoretically neutral machines in practice cannot fully avoid the impact of biases, human oversight should be included to mutually and jointly facilitate fair treatment.

Autonomy as a principle of AI ethics refers to the user's power to self-decide and whether to decide at all (Floridi et al., 2018)

Explicability

While beneficence, non-maleficence, autonomy and justice are 4 core principles in bioethics (Beauchamp & Childress, 2001), explicability was added specifically for the ethical use of AI (Floridi et al., 2018). The disclosure of more explanatory information can further diminish fairness concerns. AI decisions, particularly in black-box systems, tend to be untransparent or even incomprehensible for both qualified and non-experienced AI users. A lack of causal information for explanation can result in the misinterpretation of initial rationales leading users to perceive fairness misconceptions. Such a problem has arisen with the development and application of a Covid-19 vaccine distribution algorithm in the Stanford Medical Center (Guo & Hao, 2020). A series of protests ensued when the majority of resident physicians assigned to Covid-19 control found themselves in vaccination prioritization behind administrators or doctors who saw patients mainly remotely. An expansion of transparency and the resulting opportunity to obtain more concrete explanations might eliminate fairness issues, as users can estimate the discriminatory nature of treatments more appropriately. At the same time, transparency can reveal potential roots of unfairness since latent reasons are better detectable. Therefore, explicability can be a suitable addition in pursuing the goal of fair algorithms.

Increasing the algorithm's transparency, however, might incorporate trade-offs. The disclosure of causal reasoning might simultaneously unveil sensitive information and inherent connections in the dataset. The stigmatization of certain groups might be the result. One example is the outbreak of the HI-Virus in the early 1980s. Initially wrongly assuming that the disease primarily affects homosexual men, massive exclusion and hostility against the gay community were the result (Robert Koch-Institut, 2017). Although it was not AI that suggested the problematic inference at that time, ambiguous and questionable connections might be increasingly proposed with a growth of more systematic data analysis. Therefore, accuracy of proposed relations, as well as a careful investigation and evaluation of the information considered to be disclosed are needed to improve fairness, but at the same time avoid harm.

The purpose of explicability is to empower the other four fundamentals with a suitable amount of transparency creating more accountability and intelligibility especially for non-experienced AI users

Final Thoughts

Targeting multiple ethical principles and the goal of simultaneously optimizing all of them is clearly a challenge. Various interests of different population groups, such as the right of freedom and autonomy vs. the right to be protected against harm, need to be considered and weighed. Especially health surveillance technology requires a careful ethical investigation, as it impacts broader society and their basic needs. The European Commission has acknowledged the need for further legislation to create harmonized rules for the use of AI (European Commission, 2021). Taking a risk assessment and management approach, the European Commission even aims to prohibit certain uses of AI, such as social scoring or indiscriminate surveillance. While this is a step towards a more ethical use of AI, especially considering fairness and discrimination concerns, the EU AI Act is sometimes criticized for not going far enough (MacCarthy & Propp, 2021). One reason might be that agreeing on clear definitions of highly subjective concerns, such as fairness, and their implementation into technology is challenging. A recent example highlighting such obstacles is the algorithm developed for supporting Covid-19 vaccine distribution in the US. While it has been deployed to simplify and coordinate vaccine dissemination activities, considerable disparities in vaccine access were the result of a disharmonized priority setting in terms of medicine allocation (Singer, 2021). Oregon, for example, prioritized teachers, while New Jersey rated smokers higher in the risk of suffering from Covid-19 (Singer, 2021). It depicts one demanding issue with fairness, as agreeing on a common definition is challenging from a social point of view. Therefore, its implementation into technical measures is an even more complex endeavor. More interdisciplinary research is needed that engages both social and technical sciences to jointly work towards the common goal of serving the well-being of the public and increasing preparedness for future health crises.

References

Aiello, A. E., Renson, A., & Zivich, P. N. (2020). Social Media–and Internet-Based Disease Surveillance for Public Health. *Annual Review of Public Health*, 41, 101-118.

Apple. (2020). *Privacy-Preserving Contact Tracing*. Apple. https://covid19.apple.com/contacttracing

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics*. Oxford University Press, USA.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., & Mojsilovic, A. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv repreprint arXiv:1810.01943.*

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American journal of epidemiology*, 188(12), 2222-2239.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32.*

Budd, J., Miller, B. S., Manning, E. M., Lampos, V., Zhuang, M., Edelstein, M., Rees, G., Emery, V. C., Stevens, M. M., & Keegan, N. (2020). Digital technologies in the public-health response to COVID-19. *Nature medicine*, 1-10.

Cameron, E. E., Nuzzo, J. B., Bell, J. A., Nalabandian, M., O'Brien, J., League, A., Ravi, S., Meyer, D., Snyder, M., Mullen, L., & Warmbrod, L. (2019). *Global Health Security Index: Building Collective Action and Accountability.* https://www.ghsindex.org/wp-content/uploads/2019/10/2019-Global-Health-Security-Index.pdf

Chiolero, A., & Buckeridge, D. (2020). Glossary for public health surveillance in the age of data science. *J Epidemiol Community Health*, 74(7), 612-616.

Choi, B. C. (2012). The past, present, and future of public health surveillance. *Scientifica*, 2012

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810.*

Commission Recommendation (EU). (2020). *2020/518 of 8 April 2020 on a common Union toolbox for the use of technology and data to combat and exit from the COVID-19 crisis, in particular concerning mobile applications and the use of anonymised mobility data.* https://op.europa.eu/en/publication-detail/-/publication/1e8b1520-7e0c-11ea-aea8-01aa75ed71a1/language-en

European Commission. (2021, April 21). *Regulation of the European Parliament and of the Council: Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts.*

Fairchild, A. L., Haghdoost, A. A., Bayer, R., Selgelid, M. J., Dawson, A., Saxena, A., & Reis, A. (2017). Ethics of public health surveillance: new guidelines. *The Lancet Public Health*, 2(8), e348-e349.

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI: Challenges and Opportunities. *Business & information systems engineering*, 62(4), 379-384.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., & Rossi, F. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707.

Floridi, L., Luetge, C., Pagallo, U., Schafer, B., Valcke, P., Vayena, E., Addison, J., Hughes, N., Lea, N., & Sage, C. (2019). Key ethical challenges in the European medical information framework. *Minds and Machines*, 29(3), 355-371.

Gasser, U., Ienca, M., Scheibner, J., Sleigh, J., & Vayena, E. (2020). Digital tools against COVID-19: taxonomy, ethical challenges, and navigation aid. *The Lancet Digital Health*.

Gerke, S., Shachar, C., Chai, P. R., & Cohen, I. G. (2020). Regulatory, safety, and privacy concerns of home monitoring technologies during COVID-19. *Nature medicine*, 26(8), 1176-1182.

Gerstman, B. B. (2013). *Epidemiology kept simple: an introduction to traditional and modern epidemiology.* John Wiley & Sons.

GSMA. (2021). *The Mobile Economy 2021*. GSMA. https://www.gsma.com/mobileeconomy/wp-content/uploads/2021/07/GSMA_MobileEconomy2021_3.pdf

Guo, E., & Hao, K. (2020, December 21). *This is the Stanford vaccine algorithm that left out frontline doctors*. MIT Technology Review. https://www.technologyreview.com/2020/12/21/1015303/stanford-vaccine-algorithm/

Haar, J. M., & Spell, C. S. (2009). How does distributive justice affect work attitudes? The moderating effects of autonomy. *The International Journal of Human Resource Management*, 20(8), 1827-1842.

Klingler, C., Silva, D. S., Schuermann, C., Reis, A. A., Saxena, A., & Strech, D. (2017). Ethical issues in public health surveillance: a systematic qualitative review. *BMC Public Health*, 17(1), 295.

Laaff, M. (2020, July 3*). Corona-Warn-App – App trifft Amt.* Zeit Online. https://www.zeit.de/digital/2020-07/corona-warn-app-gesundheitsamt-hotline-labor-anbindung/komplettansicht

Lee, L. M. (2019). Public Health Surveillance: Ethical Considerations. *The Oxford Handbook of Public Health Ethics*, 320.

Lee, L. M., & Thacker, S. B. (2011). Public health surveillance and knowing about health in the context of growing sources of health data. *American journal of preventive medicine*, 41(6), 636-640.

Lee, L. M., Thacker, S. B., & Louis, M. E. S. (2010). *Principles and practice of public health surveillance.* Oxford University Press, USA.

Lucivero, F., Hallowell, N., Johnson, S., Prainsack, B., Samuel, G., & Sharon, T. (2020). Covid-19 and Contact Tracing Apps: Ethical challenges for a social experiment on a global scale. *Journal of bioethical inquiry*, 1-5.

MacCarthy, M., & Propp, K. (2021, May 4). *Machines learn that Brussels writes the rules: The EU's new AI regulation*. Brookings. https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/

Marckmann, G., Schmidt, H., Sofaer, N., & Strech, D. (2015). Putting Public Health Ethics into Practice: A Systematic Framework [Methods]. *Frontiers in Public Health*, *3*(23). https://doi.org/10.3389/fpubh.2015.00023

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Mello, M. M., & Wang, C. J. (2020). Ethics and governance for digital disease surveillance. *Science*, *368*(6494), 951-954.

Mester, T. (2017, August 21). *Statistical Bias Types explained*. Data 36. https://data36.com/statistical-bias-types-explained/

Morley, J., Cowls, J., Taddeo, M., & Floridi, L. (2020). Ethical guidelines for COVID-19 tracing apps. In: *Nature Publishing Group*.

Murff, H. J., FitzHenry, F., Matheny, M. E., Gentry, N., Kotter, K. L., Crimin, K., Dittus, R. S., Rosen, A. K., Elkin, P. L., & Brown, S. H. (2011). Automated identification of postoperative complications within an electronic medical record using natural language processing. *Jama*, *306*(8), 848-855.

Neill, D. B. (2012). New directions in artificial intelligence for public health surveillance. *IEEE Intelligent Systems*, *27*(1), 56-59.

Nuffield Council on Bioethics. (2020). Ethical Considerations in Responding to the COVID-19 Pandemic. *Nuffield Council on Bioethics, 2020*. www.nuffieldbioethics.org/assets/pdfs/Ethical-considerations-in-responding-to-the-COVID-19-pandemic.pdf

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13. https://www.frontiersin.org/articles/10.3389/fdata.2019.00013/full

Robert Koch-Institut. (2017, October 16). *1981 bis 1990: AIDS – die politische Dimension in den 1980er Jahren*. Robert Koch-Institut. https://www.rki.de/DE/Content/Institut/Geschichte/Bildband_Salon/1981-1990.html

Robert Koch-Institut. (2021). *Infektionsketten digital unterbrechen mit der Corona-Warn-App – Die Corona-Warn-App ist ein wichtiger Baustein der Pandemiebekämpfung*. Robert Koch-Institut. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/WarnApp/Warn_App.html

Rocklöv, J., Tozan, Y., Ramadona, A., Sewe, M. O., Sudre, B., Garrido, J., de Saint Lary, C. B., Lohr, W., & Semenza, J. C. (2019). Using big data to monitor the introduction and spread of Chikungunya, Europe, 2017. *Emerging Infectious Diseases*, *25*(6), 1041. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6537727/

Singer, N. (2021, February 7). *Where Do Vaccine Doses Go, and Who Gets Them? The Algorithms Decide*. The New York Times. https://www.nytimes.com/2021/02/07/technology/vaccine-algorithms.html

Statista. (2021, April). *Distribution of Twitter users worldwide as of April 2021, by age group*. Statista. https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/

Teutsch, S. M., & Churchill, R. E. (2000). *Principles and practice of public health surveillance*. Oxford University Press, USA.

Thomson, M. C., Doblas-Reyes, F., Mason, S. J., Hagedorn, R., Connor, S. J., Phindela, T., Morse, A., & Palmer, T. (2006). Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature*, *439*(7076), 576-579.

TUM Institute for Ethics in AI. (2020). *Ethical Implications of the Use of AI to Manage the COVID-19 Outbreak* (No. 2). https://doi.org/https://ieai.mcts.tum.de/wp-content/uploads/2020/04/April-2020-IEAI-Research-Brief_Covid-19-FINAL.pdf

World Health Organization. (2017). *WHO guidelines on ethical issues in public health surveillance.* https://apps.who.int/iris/bitstream/handle/10665/255721/9789241512657-eng.pdf;jsessionid=3773A3C711BBBFE02E2B623EC46D572D? sequence=1

World Health Organization. (2020, May 28). *Ethical considerations to guide the use of digital proximity tracking technologies for COVID-19 contact tracing: interim guidance.* https://www.who.int/publications/i/item/WHO-2019-nCoV-Ethics_Contact_tracing_apps-2020.1

World Health Organization. (2021a). *About WHO*. World Health Organization. https://www.who.int/about

World Health Organization. (2021b). *Ethics and Governance of Artificial Intelligence for Health*. https://apps.who.int/iris/bitstream/handle/10665/341996/9789240029200-eng.pdf

World Health Organization. (2021c). *Surveillance in emergencies*. World Health Organization. https://www.who.int/emergencies/surveillance