frontiers in Big Data

MINI REVIEW
published: 02 September 2020
doi: 10.3389/fdata.2020.00025

Considerations for a More Ethical Approach to Data in AI: On Data Representation and Infrastructure

Alice Baird[1]* *and Björn Schuller*[1,2]

[1] *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany,* [2] *Group on Language, Audio & Music, Imperial College London, London, United Kingdom*

Data shapes the development of Artificial Intelligence (AI) as we currently know it, and for many years centralized networking infrastructures have dominated both the sourcing and subsequent use of such data. Research suggests that centralized approaches result in poor representation, and as AI is now integrated more in daily life, there is a need for efforts to improve on this. The AI research community has begun to explore managing data infrastructures more democratically, finding that decentralized networking allows for more transparency which can alleviate core ethical concerns, such as selection-bias. With this in mind, herein, we present a mini-survey framed around data representation and data infrastructures in AI. We outline four key considerations (*auditing*, *benchmarking*, *confidence and trust*, *explainability and interpretability*) as they pertain to data-driven AI, and propose that reflection of them, along with improved interdisciplinary discussion may aid the mitigation of data-based AI ethical concerns, and ultimately improve individual wellbeing when interacting with AI.

Keywords: artificial intelligence, machine learning, ethical AI, decentralization, selection-bias

OPEN ACCESS

Edited by:
Fabrizio Riguzzi,
University of Ferrara, Italy

Reviewed by:
Stefania Costantini,
University of L'Aquila, Italy
Radu Prodan,
Alpen-Adria-Universität Klagenfurt, Austria

***Correspondence:**
Alice Baird
alicebaird@ieee.org

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Big Data

Received: *16 January 2020*
Accepted: *02 July 2020*
Published: *02 September 2020*

Citation:
Baird A and Schuller B (2020) Considerations for a More Ethical Approach to Data in AI: On Data Representation and Infrastructure. Front. Big Data 3:25. doi: 10.3389/fdata.2020.00025

1. INTRODUCTION

Artificial intelligence (AI) in its current form relies heavily on large quantities of data (Yavuz, 2019), and data-driven Deep Neural Networks (DNNs) have prompted fast-paced development of AI (Greene, 2020). Currently, the research community is under great strain to keep up with the potential ethical concerns which arise as a result of this (Naughton, 2019). Within the AI community such ethical concerns can require quite some disentanglement (Allen et al., 2006), and it is not until recently that AI-based research groups have begun to provide public manifestos concerning the ethics of AI, e.g., Google's DeepMind, and the Partnership AI.[1]

The *Ethics of AI* (Boddington, 2017) is now an essential topic for researchers, both internal and external, to core-machine learning and differs from *Machine Ethics* (Baum et al., 2018). The latter refers to giving conscious ethical based decision-making power to machines. The *Ethics of AI*, although somewhat informing *Machine Ethics*, refers more broadly to decisions made by researchers and covers issues of diversity and representation, e.g., to avoid discrimination (Zliobaite, 2015) or inherent latent biases (van Otterlo, 2018). Herein, our discussion focuses on topics relating to the *Ethics of AI* unless otherwise stated.

There has been recent research which shows promise for improved data learning from smaller quantities ("merely a few minutes") of data (Chen et al., 2018). However, machine learning

[1] DeepMind : https://deepmind.com/applied/deepmind-ethics-society/. Partnership on AI: https://www.partnershiponai.org/board-of-directors/.

algorithms developed for AI commonly require substantial quantities of data (Schneider, 2020). In this regard, *Big Data* ethics for AI algorithms are an expanding discussion point (Berendt et al., 2015; Mittelstadt and Floridi, 2016). Crowdsourcing (i.e., data gathered from large amounts of paid or unpaid individuals via the internet), is one approach to collect such quantities of data. However, ethical concerns including worker exploitation (Schlagwein et al., 2019), may have implications on the validity of the data. Additionally researchers utilize *in-the-wild* internet sources, e.g., YouTube (Abu-El-Haija et al., 2016) or Twitter (Beach, 2019), and apply unsupervised labeling methods (Jan, 2020). However, in Parikh et al. (2019), the authors describe how approaches for automated collection and labeling can result in the propagation of historical and social biases (Osoba and Welser IV, 2017). In the health domain, such bias could have serious consequences, leading to misdiagnosis or incorrect treatment plans (Mehrabi et al., 2019).

One method to avoid bias in AI is through the acquisition of diverse data sources (Demchenko et al., 2013). With *Veracity* (i.e., habitual truthfulness) being one of the *5 Vs* (e.g., Velocity, Volume, Value, Variety and Veracity) for defining truly Big Data (Khan et al., 2019). However, big data is commonly, stored in *centralized* infrastructures which limit transparency, and democratic, decentralized (i.e., peer-to-peer blockchain-based) approaches are becoming prevalent (Luo et al., 2019).

Centralized data storage can be efficient and beneficial to the "central" body to which the infrastructure belongs. However, it is precisely this factor amongst others (i.e., proprietary modeling of underrepresented data) that are problematic (Ferrer et al., 2019).

Furthermore, centralized platforms limit the access and knowledge that data providers receive. The General Data Protection Regulation (GDPR) was established within the European Union (The-European-Commission, 2019) to partly tackle this. GDPR is a set of regulations of which the core goal is to protect the data of individuals that are utilized by third parties. In its current form, GDPR promotes a centralized approach, supporting what are known as *commercial governance platforms*. These platforms control restrictions to employees based on a data providers request but primarily function as a centralized repository. In essence, GDPR meant that companies needed to re-ask for data-consent more transparently. However, the "terms of agreement" certificate remains the basis, and 90 % of users are known to ignore its detail (Deloitte, 2016).

As a counter approach to the centralized storage of data, for some time researchers have proposed the need for a *decentralized* (cf. **Figure 1**) networking in which individual data is more easily protected (i.e., there is no "single point" of failure). In this infrastructure, individuals have more agency concerning the use of their data (Kahani and Beadle, 1997). Primarily, individuals choose to access parts of a network rather than its entirety. On a large scale, this paradigm would remove the known biases of centralized networks, as targeted collection, for example, would be less accessible by companies and sources of the data more complex to identify. In this way, various encryption algorithms, including homomorphic encryption (a method which allows for data processing while encrypted), or data masking, are being integrated within decentralized networks, allowing for identity preservation (Setia et al., 2019). Federated Learning (FL) (Hu et al., 2019), is one approach which can be applied to decentralized networks to improve privacy (Marnau, 2019). In FL, weights are passed from the host device and updated locally, instead of raw data leaving a device (Yang et al., 2019).

With these topics in mind, in this contribution, we aim to outline core ethical considerations, which relate to data and the ethics of AI. Our focus remains on the ethics of data representation and data infrastructure, particularly *selection-bias* and *decentralization*. We chose these topics due to their common pairing in the literature. A regular talking-point in machine learning is *selection-bias* and a networking infrastructure which may help to more transparently observe this is *decentralization* (Swan, 2015; Montes and Goertzel, 2019).

Our contribution is structured as follows; firstly we shortly define key terminology used throughout the manuscript in section 2, followed by a brief background and overview of the core themes as they pertain to AI in section 3. We then introduce our ethical data considerations in section 4 providing specific definitions and general ethical concerns. Following this in section 5, we connect these ethical considerations more closely with data representation and infrastructure, and in turn, outline technical approaches which help reduce the aforementioned ethical concerns. Finally, we offer concluding remarks in section 7.

2. TERMINOLOGY

There are a variety of core terms which are used throughout this manuscript which may have a dual meaning in the machine learning community. For this reason, we first define here three core terms, *ethics*, *bias*, and *decentralization* used within our discussion.

As mentioned previously, we focus on the *Ethics of AI* rather than *Machine Ethics*. However, further to this, we use the term *ethics* based on guidelines within applied ethics, particularly in relation to machine understanding. In Döring et al. (2011), the principles of *beneficence*, *non-maleficence*, *autonomy*, and *justice* are set out as being fundamental considerations for those working in AI. Although this is particular to emotionally aware systems, we consider that such principles are relevant across AI research. Of particular relevance to this contribution, is autonomy, i.e., a duty for systems to avoid interference, and respect an individual's capacity for decision-making. This principle impacts upon both *data representation* and *infrastructure* choices (e.g., centralized or decentralized).

We consistently refer to the term *bias* throughout our contribution. First introduced to machine learning by Mitchell (1980), we typically discuss statistical biases, unless otherwise stated, which may include absolute or relative biases. To be more specific, we focus closely on data in this contribution, and therefore dominantly refer to *selection-bias*. *Selection-bias* stems in part from prejudice-based biases (Stark, 2015). However, *selection-bias* falls within statistical biases as it is a consequence of conscious (hence prejudice) or unconscious data

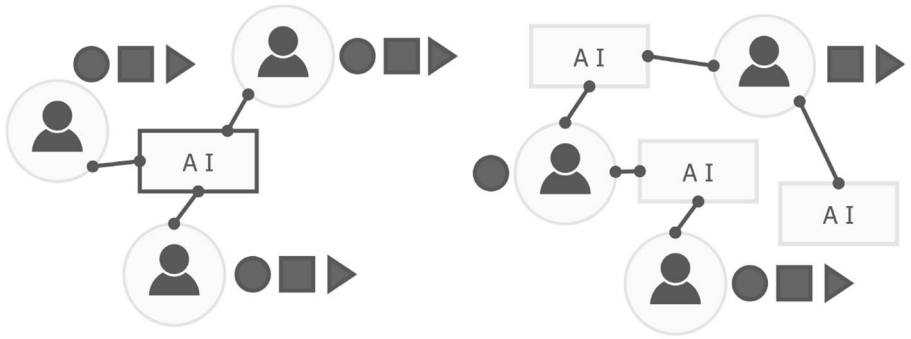

FIGURE 1 | A simplified overview of a typical centralized (left) and decentralized (right) network infrastructure. In the right figure individuals choose the modality to share (as indicated by circle, square, and triangle icons), and users in the network have agency in how their data is used. In the left figure, the AI is essentially a black-box, and users make all modalities of data available to all components of the AI infrastructure.

selection. *Selection-bias* is particularly relevant to AI given that real randomization (or diverse representation) of data is not always possible.

As a critical aspect of our contribution, relating to the mitigation of bias, through a more ethical approach to data infrustructure, we consistently refer to *decentralized* AI. A broad definition of *decentralization* is the distribution of power moving away from central authorities. In the context of AI, when discussing *decentralization*, we refer to decentralized architectures which allow for this type of distribution, in regards to data sourcing, management and analysis. We do touch on literature relating to blockchain, which is a well-known decentralized approach. However, the term is utilized here more generally and is not exclusive to the blockchain.

3. BACKGROUND: BIAS AND DECENTRALIZATION IN AI

Funding and global research efforts in the field of AI have increased in the last decade, particularly in the areas of health, transportation, and communication (Mou, 2019). Along with this increase has come a rise in ethical demands related to Big Data (Herschel and Miori, 2017). Although *true* Big Data is said to need *Veracity*, the reality of this is sometimes different, with large-scale data often showing particular biases toward clustered demographics (Price and Ball, 2014). As a result, terms, such as *Machine Learning Fairness*—promoted initially by Google Inc.[2]—is now regularly referred to in an endeavor to build *trust* and show ethical sensitivity (Mehrabi et al., 2019). In this regard, IBM released their AI Explainability 360 Toolkit[3] in which the overarching goal appears to be improving *trust* in AI, through more deeply researching machine learning biases, as it pertains to the research areas of fairness, robustness and *explainability*.

Three common forms of bias are discussed concerning AI, i.e., interaction-bias, latent-bias, and *selection-bias*. *Selection-bias* occurs when the data used within a paradigm is selected with bias, leading to misrepresentation rather than generalization. In particular, researchers are repeatedly finding bias in regards to gender (Gao and Ai, 2009). Wang et al. (2019a) found for example that models tend to have a bias toward a particular gender even when a dataset is balanced—which could point to lower level architecture-based biases (Koene, 2017). *Selection-bias* is essential to combat when referring to models developed for human interaction. Based on data decision making, a bias can propagate through system architectures, leading to lower accuracy on a generalized population. Lack of generalization is particularly problematic for domains, such as health, where this may result in a breach of patient safety (Challen et al., 2019).

Furthermore, the evaluation of *fairness* in machine learning is another prominent topic, highlighted as a machine learning consideration in Hutchinson and Mitchell (2019). Additionally, researchers propose *fairness metrics* for evaluating the bias which is inherent to a model (Friedler et al., 2019), including the Disparate Impact or Demographic Parity Constraint (DPC). DPC groups underprivileged classes and compares them to privileged classes as a single group. Similarly, there are novel architectures which mitigate bias through prioritization of minority samples, and the authors of this approach suggest that there is an improvement in *generalized fairness* (Lohia et al., 2019).

A core contributing factor to bias in AI is the management of data. Current AI networking is based on centralized infrastructure (cf. **Figure 1**), where individuals present a unified data source to a central server. This centralization approach not only limits privacy but also creates a homogeneous representation, which is less characteristic of the individual interacting (Sueur et al., 2012).

Decentralization in AI was initially coined as a term to describe "autonomous agents in a multi-agents world" (Miiller, 1990), and researchers have proposed *decentralization* for large AI architectures e.g., integrating machine learning with a Peer-to-peer style blockchain approach Zheng et al., 2018] to improve *fairness* and *bias* (Barclay et al., 2018). In this architecture, collaborative incentives are offered to the

[2]Google: https://developers.google.com/machine-learning/fairness-overview/.
[3]IBM AI Explainability 360 Toolkit: https://www.research.ibm.com/artificial-intelligence/trusted-ai/.

network users and approaches allow for improved identity-representation, as well as more control in regards to data-usage, resulting in more freedom and higher privacy. Furthermore, a decentralized network may inherently be more ethical as more individuals are interacting with and refining the network with agency (Montes and Goertzel, 2019).

For individuals interfacing with AI, privacy is a concern (Montes and Goertzel, 2019). Improving privacy is a core advantage of decentralized data approaches (Daneshgar et al., 2019). In a centralized approach, anonymization processes exist (e.g., that which are enforced by GDPR), although it is unclear how this is consistently applied. To this end, identification of a participant in the data source may not be needed, yet, unique aspects of their character (e.g., how they pronounce a particular word), are still easily identified (Regan and Jesse, 2019).

There are multiple organizations and corporations which focus on the benefits of *decentralization*, including Effect.AI and SingularityNET[4] Such organizations promote benefits including "diverse ecosystems" and "knowledge sharing." The Decentralized AI Alliance[5] is another organization which integrates AI and blockchain, promoting collaborative problem-solving. In general, the term *decentralization* comes not only from technical network logistic but from philosophical "transhuman" ideologies (Smith, 2019). In regards to the latter, *decentralization* promotes the improvement of human-wellbeing through democratical interfacing with technology Goertzel (2007). This democratic view is one aspect of *decentralization* that aids in the reduction of AI bias (Singh, 2018).

Similarly, there are organizations which focus primarily on the challenge of bias in AI, from many viewpoints including race, gender, age, and disability[6], most of which implement responsible research and innovation (RRI). When applying RRI to the AI community, the aim is to encourage researchers to anticipate and analyse potential risks of their network, and ensure that the development of AI is socially acceptable, needed, and sustainable (Stahl and Wright, 2018). Biases are an essential aspect of AI RRI (Fussel, 2017), as poor identity-representation has dire consequences for real-world models (Zliobaite, 2015).

4. METHODOLOGY: ETHICAL DATA CONSIDERATIONS

There are an array of concerns relating to the ethics of AI, including, joblessness, inequality, security, and prejudices (Hagendorff, 2019). With this in mind, academic and industry-based research groups are providing tools to tackle these ethical concerns (cf. **Table 1**), mainly based on four key areas. In this section, we introduce and conceptually discuss these four ethical considerations—*auditing*, *benchmarking*, *confidence and trust* and *explainability and interpretability*—chosen, due to their prominence within the AI community. As well as this, these four aspects, each have a pivotal impact on data representation, and an inherent relation to data infrastructures. An overview of a typical machine learning workflow with these four considerations highlighted based on their position in time is given in **Figure 2**. To this end, herein, we first define our four considerations more concretely, followed by a description of specific ethical concerns ([±]) which relate to them.

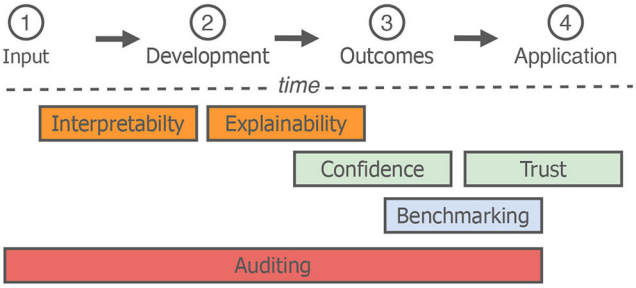

FIGURE 2 | An overview of a machine learning workflow, (1) Data collection and pre-processing, (2) developments of machine learning models, (3) evaluation of model outcomes (i.e., performance), (4) integration of the developed AI in a real-world scenario. We place the four considerations introduced in Section 4 across time. Positions of individual considerations are not static, we define their placement over time, based primarily on their relationship to one another.

4.1. Auditing

In the context of AI data, *auditing* is not dissimilar to research domains, such as economics. An auditor regularly checks aspects of the system, including the data validity itself. For example Fernández and Fernández (2019) propose an AI-based recruiting systems—in which the candidate's data is validated by a manual (i.e., human) auditor. In **Figure 2** we have assigned *auditing* to every aspect of the AI workflow, although it is commonly only integrated during earlier development stages.

[±] *Auditing* is integral as acquisition scales up to *Big Data*. The process of managing what Schembera and Durán (2020), describes as "tangible data" can be extremely time-consuming and costly for those involved and human or machine error can propagate, resulting in biases or leading to mostly unusable data (L'heureux et al., 2017). On the other side, is the *auditing* of "dark data." This data type is estimated to be 90% (Johnson, 2015) of all stored data, and is largely unknown to the user. The literature currently focuses on *auditing* tangible data, as yet there is less attention for dark data (Trajanov et al., 2018).

4.2. Benchmarking

In machine learning, *benchmarking* is the process of evaluating novel approaches against well-establish approaches or databases of the same task. To this end, it often comes at a later stage during the AI workflow (cf. **Figure 2**). In the computer vision domain, this has been particularly successful in pushing forward developments (Westphal et al., 2019), with data sets, such as MNIST (LeCun and Cortes, 2010) or CIFAR-10 (Krizhevsky et al., 2009), continuously benchmarked against in both an

[4]Effect.AI: https://effect.ai/., SingularityNET https://singularitynet.io/.
[5]Decentralized AI Alliance: https://daia.foundation/.
[6]The Algorithm Justice League: https://www.ajlunited.org/, and the AI NOW institute https://ainowinstitute.org/.

TABLE 1 | Brief overview of prominent ethical AI tools which have been made available by both academic and industry research groups.

Tool	A	B	E & I	C & T	Description
Gender Shades (Buolamwini and Gebru, 2018)	X	X	–	–	An *intersectional* approach to inclusive product testing for AI, relating specifically to gender and race bias.
What-If Tool (Google, 2020)	X	–	X	–	Allows users to analyse their machine learning model through the use of an interactive visual interface.
IBM: AI Explainability 360 Toolkit (Arya et al., 2020)	–	X	X	–	Contains state-of-the-art algorithms that allow for improved interpretability and explainability of machine learning models.
IBM: AI Fairness 360 Open Source Toolkit (Bellamy et al., 2019)	X	–	X	X	Provides a series of metrics for datasets and models to test for biases explicitly, including a clear explanations for those metrics.
LIME (Ribeiro et al., 2016)	–	–	X	X	A general eXplainable-AI toolkit which allows users to reason better for why a model makes certain predictions.
openAI: baseline, Gym, Microscope (Brockman et al., 2016)	–	X	X	–	Provides reproducible reinforcement learning algorithms with benchmarked performances based on published results. As well as visualization methods for observing significant layers and neuron activations.
Procgen: Benchmark (Cobbe et al., 2019)	–	X	–	–	Procedurally-generated environments which provide a benchmark for the speed of a reinforcement learning algorithms generalization.
PwC: Responsible AI Toolkit (Waterhouse Cooper, 2019)	–	–	X	X	A collection of customizable frameworks to harness AI in an ethical and responsible manner.
Pymetrics: Audit AI (Trindel et al., 2019)	X	–	–	–	Contains tools to measure and mitigate the effects of discriminatory patterns, designed specifically for socially sensitive decision processes.

We highlight their target ethical consideration, namely (A)uditing, (B)enchmarking, (E)xplainability and (I)nterpretabiltiy, (C)onfidence and (T)rust.

academic and industry setting. Pre-trained networks are another *benchmarking* tool. Networks, such as imageNet (Simon et al., 2016) are well-known and consistently applied, given the quantity of data and promising results (Wang et al., 2019d).

[±] Multimodal analysis is becoming more ubiquitous in machine learning (Stappen et al., 2020), due to well-known and longstanding advantages (Johnston et al., 1997). When datasets are multimodal *benchmarking* improvements accurately becomes complex (Liu et al., 2017), and aspects, such as modality miss-matches are common (Zhang and Hua, 2015). Additionally, given the rapid developments in machine learning approaches, outdated methods may be held as benchmarks for longer than is scientifically meaningful.

4.3. Confidence and Trust

In AI data, the terms *confidence and trust* are applied to ensure reliability, i.e., having *confidence* in the data results in deeper *trust* (Arnold et al., 2019). In this context, *trust* is a qualitative term, and although *confidence* can fall into these interpretations relating to enhanced moral understanding (Blass, 2018), the term *confidence* typically refers to a quantifiable measure to base *trust* on (Zhang et al., 2001; Keren et al., 2018).

[±] Not providing an overall *confidence* for resulting predictions, can result in a substantial risk to the user (Ikuta et al., 2003), i.e., if a trained network has an inherent bias, a *confidence* measure improve the transparency of this. Furthermore, to increase *trust* in AI, developers are attempting to replicate human-like characteristics, e.g., how robots walk (Nikolova et al., 2018). Adequately reproducing such characteristics, requires substantial data sources from refined demographics. This concern falls primarily into *Machine Ethics*, with the need for binary gender identifications (Baird et al., 2017), and the societal effect of doing so challenged (Jørgensen et al., 2018).

4.4. Explainability and Interpretability

Often referred to as XAI (eXplainable AI) and arguably at the core of the ethical debate in the field of AI is *explanabilty* and *interpretability*. These terms are synonymous for the need to understand algorithms' decision making (Molnar, 2019; Tjoa and Guan, 2019). However, a distinction can be made, *interpretability* being methods for better understanding a machine learning architecture or data source (i.e., the *how*), and *explainability* being methods for understanding *why* particular decision were made.

[±] A surge in machine learning research, has come from international challenges (Schuller et al., 2013; Ringeval et al., 2019)—driving improvements in accuracy across multiple machine learning domains (Meer et al., 2000). However, this fast-paced environment often leaves less time for interpreting how particular features may have explicitly impacted a result, or for an explanation of a models decision-making process. Without this, the meaning of any result is less easy to substantiate (Vellido et al., 2012).

5. DISCUSSION: REPRESENTATION AND INFRASTRUCTURE

Having defined our four key consideration more concretely, we now discuss them more closely with representation (w.r.t., bias) and infrastructure of AI data in mind. Where meaningful, we highlight technical approaches which are implemented to reduce the aforementioned ethical concerns.

5.1. Auditing

There are many methods being developed to make collecting and annotating data in an automatic way possible, including *data mining* of web-based images (Zafar et al., 2019), and *active learning* (AL) for semi-automatic labeling (Wang et al., 2019c). For data tagging by autonomous agents, some have shown concerns that making agents responsible for this, may lead to incorrect tagging caused by an initial human error. A concern which becomes more problematic given the now large quantities of child viewers, who may be *suggested* inappropriate content (Papadamou et al., 2019). Further to this when annotating data, one ethical issue which can propagate *selection-bias* is poorly balanced manual vs. automatic annotations. In other words, if automatic annotation procedures learn false aspects early on, these may then be replicated (Rothwell et al., 2015). In an AL paradigm (Ayache and Quénot, 2008), an *oracle* (i.e., expert auditor) is kept in the loop, and where the AL model is uncertain at a particular level of *confidence*, the oracle must provide the label (Settles et al., 2008). In the case of specialist domains, such as bird sound classification, having such an expert is crucial, as variances in the audio signal can be quite slight (Qian et al., 2017).

Within a larger *decentralized* network, utilizing auditors allows for a democratic style of data management. Blockchain AI networks, for example, run in a peer-to-peer (P2P) fashion, meaning that no changes can be made to the system without the agreement of all others in the network. In a P2P network, there is an incentive for individual participation in the *auditing* process (e.g., an improved overall experience) (Dinh and Thai, 2018). However, the realization of *auditing* in AI does lead to some technical challenges in regards to public verification of sensitive data (Diakopoulos and Friedler, 2017), as well as making the AI only a partial reduction of human time-cost. Nevertheless, the need for *auditing* in AI has been highlighted consistently in the literature as a bias mitigating approach (Saleiro et al., 2018)

5.2. Benchmarking

It has been noted in many domains of research that *benchmarking* and therefore generalizing against a well-established organization, may result in the continued propagation of poor standards concerning historical biases (Denrell, 2005). Survey-based evaluations of the state-of-the-art modalities and baselines results are one resource to help mitigate this issue (Liu et al., 2011; Cummins et al., 2018). However, constant updates to benchmarks should be made, updating both techniques for acquisition and methods for setting baselines. Although there is no rule of thumb in this case, it is generally accepted in machine learning that *benchmarking* against resources that are no longer considered to be state-of-the-art will not bring valid results. Furthermore, in the realm of human-data, and specifically within the European Union, there is often a limited time that data can be stored (The-European-Commission, 2019). In this way, not only will benchmarked data sets become outdated in terms of techniques, but it is unethical to utilize such data, as reproducibility may not be possible.

Of note, a considerable contribution for ethics-based *benchmarking* is the aforementioned open-source IBM AI Explainability 360 Toolkit, in which one aspect is the Adversarial Robustness 360 Toolbox. This toolbox provides state-of-the-art paradigms for adversarial attacks (i.e., subtle alterations to data), and allows researchers to benchmark their approaches in a controlled environment to allow for more easy *interpretation* of possible network issues.

5.3. Confidence and Trust

Given the general fear that members of the public have for AI—mostly attributed to false depictions in movies and literature – improving *confidence and trust* in AI is now at the forefront for many corporations. To this end, researchers and corporations continually introduce state-of-the-art aids for tackling famous AI problems, such as the IBM AI Fairness 360 Toolkit. As well as this, to improve *trust* groups, such as "IBM Building Trust in AI"[7], make this their specific focus. In this particular group, developing human-like aspects is given a priority, as research has shown that humans *trust* the general capability of more human-like representations over purely mechanical ones (Charalambous et al., 2016). However, the well-known uncanny valley (which refers to familiarity and likeability, concerning human-likeness) suggests that data-driven representations requiring *trust* should be very-near human-like (Mori et al., 2012), and action may result in biased binary representations, which may be problematic in terms of identity politics (Jørgensen et al., 2018).

Another effort in improving *trust* comes from blockchain. Blockchain is a specific *decentralized* approach known as a distributed digital ledger, in which transactions can only be altered with the specific agreement of subsequent (connected) blocks (Zheng et al., 2018). Blockchain is said to offer deeper *trust* for a user within a network, due to the specific need for collaboration (Mathews et al., 2017). This approach offers further

[7]IBM—Building Trust in AI: https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html.

accountability, as decisions, or alterations are agreed upon by those within the network. More specifically, *trust* is established through algorithms known as consensus algorithms (Lee, 2002).

As mentioned, one quantifiable measure to build on *trust* are *confidence measures*, sometimes referred to as *uncertainty measures* i.e., those applied in a semi-automated labeling paradigm. A *confidence* measure evaluates the accuracy of a model's predictions against a ground truth or set of weights and provides a metric of *confidence* in the resulting prediction (Jha et al., 2019). Herein, we follow this definition for *confidence* as a measure, i.e., how accurate is the current system prediction, as a means of understanding any risk (Duncan, 2015). This definition allows researchers to have a margin of error and can be a crucial aspect of the health domain to avoid false-positives (Bechar et al., 2017).

Given the "black-box" nature of deep learning, there have been numerous approaches to quantifying *confidence* (Kendall and Cipolla, 2016; Keren et al., 2018). One popular procedure for measuring *confidence* is the *Monte Carlo dropout*. In this approach, several iterations are made, each time "dropping" a portion of the network, and calculating *confidence* or uncertainty based on the variance of each prediction (Gal and Ghahramani, 2016).

As an additional note, *data-reliability* is a term often referred to in regards to both *confidence and trust*. Typically this is the process of statistically representing the significance of any findings from the database in a well-established scientific fashion, particularly considering the context of the domain it is targeted toward (Morgan and Waring, 2004). Statistical tests, such as the *p*-value, which is used across research domains, including machine learning, remains controversial. A *p*-value, states the strength (significance) of evidence provided and suffers from the "dancing *p*-value phenomena" Cumming (2013). This phenomenon essentially shows that in a more real-world setting the *p*-value can range (within the same experimental settings) from <0.001 to 0.5, i.e., from very significant to not significant all. Given this limitation, the researcher may present a biased experiment, in an endeavor to report a significant result. This limitation of the *p*-value, amongst other statistical tests, has gained criticism in recent years, due to their extensive misuse by the machine learning community (Vidgen and Yasseri, 2016).

5.4. Explainability and Interpretability

Researchers continue to work towards more accurately understanding the decisions made by deep networks (Huszár, 2015; Rai, 2020). Machine learning models must be interpretable and offer a clear use-case. At the core of this, data itself in such systems should also be explainable i.e., designed data acquisition, with plausible goals. Machine learning is a pattern recognition task, and due to this visualization of data is one way to help with detailing both *interpretability* and *explainability* of a system by (1) better understanding the feature space, and (2) better understanding possible choices. In regards to the bias in AI, visualization of data-points allows for a more easily determined observation of any class dominance. Clustering is a particular pre-processing step applied in *Big Data*-based deep learning (Samek et al., 2017). Popular algorithms which apply this type of visualization include *t*-distributed stochastic neighbor embedding (*t*-SNE) (Zeiler and Fergus, 2014) and Laplacian Eigenmaps (Schütt et al., 2019). More recently, there has been a surge in approaches for visualizing attention over data points (Guo et al., 2019). These approaches are particularly promising as they show visually the areas of activation which are learnt most consistently for each class by a network (Wang et al., 2019b), therefore highlighting areas of bias more easily, and improving communication methods to those outside the field.

To this end, *decentralization* with integrated blockchain is one approach which has been noted as improving *interpretability*, mainly as data is often-publicly accessible (Dinh and Thai, 2018). For example, where bias begins to form, the diversity of modalities and ease in identification means that individual blocks can be excluded entirely from a network to meet a more accurate representation (Dai et al., 2019).

6. FUTURE DIRECTIONS

Due in part to the ethics-based commitments by some of the larger AI companies, we see from this review that, there is momentum toward a more ethical AI future. However, **interdisciplinarity** in AI research is one aspect which requires more attention. To the best of the authors' knowledge, most public forums (particularly those based on a centralized infrastructure) come from a mono-domain viewpoint (e.g., engineering). Incorporating multiple disciplines in the discussion appears to be more prominent with those promoting *decentralized* AI.

Interdisciplinary will not only improve implementation of the four ethical consideration described herein, but has been shown to be a necessary step forward for the next AI phase of Artificial General Intelligence (AGI), proposed by the decentralized community (Goertzel and Pennachin, 2007). Interdisiplinarity is particularly of value as infrastructures developed in this way more easily tackle ethical concerns relating to; (i) integration, (ii) *selection-bias*, and (iii) *trust*.

Seamless **integration** of AI is necessary for its success and adoption by the general public. Aspects including cultural and environmental impact need to be considered, and various experts should provide knowledge on the target area. For example, the synthesized voice of bus announcements not representing the community to which it speaks may have a negative impact on those communities, and a closer analysis of the voice that best represents that community would be more ethically considerate. In this way, working alongside linguists and sociologists may aid development.

Similarly, from our literature overview, we observe that knowledge of **selection-bias** often requires contributions from experts with non-technical backgrounds, and an approach for facilitating discussion between fields of research would be a valuable next step. For example, within the machine learning community, techniques, such as *few-shot learning* are receiving more attention in recent years (Wang and Yao, 2019), however, perceptual-based biases pose difficulties for such approaches (Azad et al., 2020), and discussion from experts of

the targeted domains may help understand the bias at an earlier stage. Despite this, communication between fields speaking different "languages" (i.e., anthropology and engineering), is a challenge in itself, which should be addressed by the community. Furthermore, due to historical stereotypes, AI continues to lack in **trust** by the general user. Users who without an understanding of the vocabulary of the field, may not be able to grasp the concept of such networks. Through a better collaboration with various academic researchers, communicating AI to the general public may also see an improvement, which in turn will help to build trust and improve wellbeing of the user during AI interaction.

7. CONCLUSION

The themes of data representation and infrastructure as they pertain to *selection-bias* and *decentralization* in AI algorithms have been discussed throughout this contribution. Within these discussion points, we have highlighted four key consideration; *auditing*, *benchmarking*, *confidence and trust*, and *explainability and interpretability* to be taken into account when handling AI data more ethically.

From our observation, we conclude that for all of the four considerations, issues which may stem from multimodal approaches should be treated cautiously. In other words, relating to *auditing*, there should be standards for each modality monitored, as this follows through into the ability for accurate *benchmarking*. In this same way, although the literature may argue this, *confidence and trust* come from diverse representations of human data, which in turn are more *explainable* to the general public due to its inherent human-like attributes.

With this in mind, we see that efforts are being made, for fully audited, benchmarkable, confident, trustworthy, explainable and interpretable machine learning approaches. However, standardization for the inclusion of all of these aspects is still needed. Furthermore, with the inclusion of multiple members who take equal responsibility, *decentralization* may enable the ethical aspects highlighted herein. We see that through social-media (which is in some sense a decentralized network for communication) group morality is developed. Opinions of a political nature, for example, are highlighted, and any prejudices or general wrongdoing is often shunned and which can have enormous impact on business (Radzik et al., 2020). In this way, a more transparent and open platform makes masking potential network biases a challenge.

AUTHOR CONTRIBUTIONS

AB: literature analysis, manuscript preparation, editing, and drafting manuscript. BS: drafting manuscript and manuscript editing. All authors revised, developed, read, and approved the final manuscript.

FUNDING

This work was funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

REFERENCES

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, A. P., Toderici, G., Varadarajan, B., et al. (2016). Youtube-8m: a large-scale video classification benchmark. *arXiv* 1609.08675.

Allen, C., Wallach, W., and Smit, I. (2006). Why machine ethics? *IEEE Intell. Syst.* 21, 12–17. doi: 10.1109/MIS.2006.83

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., et al. (2019). Factsheets: increasing trust in ai services through supplier's declarations of conformity. *IBM J. Res. Dev.* 63, 6:1–6:13. doi: 10.1147/JRD.2019.2942288

Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., et al. (2020). "AI explainability 360: hands-on tutorial," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona), 696.

Ayache, S., and Quénot, G. (2008). "Video corpus annotation using active learning," in *European Conference on Information Retrieval* (Glasgow: Springer), 187–198. doi: 10.1007/978-3-540-78646-7_19

Azad, R., Fayjie, A. R., Kauffman, C., Ayed, I. B., Pedersoli, M., and Dolz, J. (2020). On the texture bias for few-shot cnn segmentation. *arXiv* 2003.04052.

Baird, A., Jørgensen, S. H., Parada-Cabaleiro, E., Hantke, S., Cummins, N., and Schuller, B. (2017). "Perception of paralinguistic traits in synthesized voices," in *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences* (London: ACM), 17. doi: 10.1145/3123514.3123528

Barclay, I., Preece, A., and Taylor, I. (2018). Defining the collective intelligence supply chain. *arXiv* 1809.09444.

Baum, K., Hermanns, H., and Speith, T. (2018). "From machine ethics to machine explainability and back," in *Proceedings of International Symposium on Artificial Intelligence and Mathematics* (Fort Lauderdale, FL: FL).

Beach, A. (2019). *Threat Detection on Twitter Using Corpus Linguistics*. Burlington, NH: University of Vermont Libraries.

Bechar, M. E. A., Settouti, N., Chikh, M. A., and Adel, M. (2017). Reinforced confidence in self-training for a semi-supervised medical data classification. *Int. J. Appl. Pattern Recogn.* 4, 107–127. doi: 10.1504/IJAPR.2017.085323

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., et al. (2019). Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* 63, 4–1. doi: 10.1147/JRD.2019.2942287

Berendt, B., Büchler, M., and Rockwell, G. (2015). Is it research or is it spying? Thinking-through ethics in big data AI and other knowledge sciences. *Kunstl. Intell.* 29, 223–232. doi: 10.1007/s13218-015-0355-2

Blass, J. A. (2018). You, me, or us: balancing individuals' and societies' moral needs and desires in autonomous systems. *AI Matters* 3, 44–51. doi: 10.1145/3175502.3175512

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Cham: Springer International Publishing.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., et al. (2016). Openai gym. *CoRR* abs/1606.01540.

Buolamwini, J., and Gebru, T. (2018). "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency* (New York, NY), 77–91.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., and Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* 28, 231–237. doi: 10.1136/bmjqs-2018-008370

Charalambous, G., Fletcher, S., and Webb, P. (2016). The development of a scale to evaluate trust in industrial human-robot collaboration. *Int. J. Soc. Robot.* 8, 193–209. doi: 10.1007/s12369-015-0333-8

Chen, Y., Assael, Y., Shillingford, B., Budden, D., Reed, S., Zen, H., et al. (2018). Sample efficient adaptive text-to-speech. *arXiv* 1809.10460.

Cobbe, K., Hesse, C., Hilton, J., and Schulman, J. (2019). Leveraging procedural generation to benchmark reinforcement learning. *arXiv* 1912.01588.

Cumming, G. (2013). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Abingdon: Routledge.

Cummins, N., Baird, A., and Schuller, B. (2018). Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. *Methods* 151, 41–54. doi: 10.1016/j.ymeth.2018.07.007

Dai, H.-N., Zheng, Z., and Zhang, Y. (2019). Blockchain for internet of things: a survey. *IEEE Internet Things J.* 6, 8076–8094. doi: 10.1109/JIOT.2019.2920987

Daneshgar, F., Sianaki, O. A., and Guruwacharya, P. (2019). "Blockchain: a research framework for data security and privacy," in *Workshops of the International Conference on Advanced Information Networking and Applications* (Caserta: Springer), 966–974. doi: 10.1007/978-3-030-15035-8_95

Deloitte, L. (2016). *Global Mobile Consumer Survey 2016*. London: Deloitte, UK Cut.

Demchenko, Y., Grosso, P., De Laat, C., and Membrey, P. (2013). "Addressing big data issues in scientific data infrastructure," in *2013 International Conference on Collaboration Technologies and Systems (CTS)* (San Diego, CA: IEEE), 48–55. doi: 10.1109/CTS.2013.6567203

Denrell, J. (2005). Selection bias and the perils of benchmarking. *Harvard Bus. Rev.* 83, 114–119. Available online at: https://hbr.org/2005/04/selection-bias-and-the-perils-of-benchmarking.

Diakopoulos, N., and Friedler, S. (2017). *How to Hold Algorithms Accountable*. MIT Technology Review. Available online at: http://bit.ly/2f8Iple

Dinh, T. N., and Thai, M. T. (2018). AI and blockchain: a disruptive integration. *Computer* 51, 48–53. doi: 10.1109/MC.2018.3620971

Döring, S., Goldie, P., and McGuinness, S. (2011). "Principalism: a method for the ethics of emotion-oriented machines," in *Emotion-Oriented Systems: The Humaine Handbook*, eds R. Cowie, C. Pelachaud, and P. Petta (Berlin; Heidelberg: Springer), 713–724. doi: 10.1007/978-3-642-15184-2_38

Duncan, B. (2015). *Importance of Confidence Intervals*. Insights Association. Available online at: http://bit.ly/2pgT4kM

Fernández, C., and Fernández, A. (2019). Ethical and legal implications of ai recruiting software. *ERCIM News* 116, 22–23. Available online at: https://ercim-news.ercim.eu/en116/special/ethical-and-legal-implications-of-ai-recruiting-software.

Ferrer, A. J., Marquès, J. M., and Jorba, J. (2019). Towards the decentralised cloud: survey on approaches and challenges for mobile, *ad hoc*, and edge computing. *ACM Comput. Surv.* 51:111. doi: 10.1145/3243929

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY: ACM), 329–338. doi: 10.1145/3287560.3287589

Fussel, S. (2017). *AI Professor Details Real-World Dangers of Algorithm Bias*. Gizmodo. Available online at: http://bit.ly/2GDoudz

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a bayesian approximation: representing model uncertainty in deep learning," in *International Conference on Machine Learning* (New York, NY), 1050–1059.

Gao, W., and Ai, H. (2009). "Face gender classification on consumer images in a multiethnic environment," in *Proceedings of International Conference on Advances in Biometrics* (Alghero), 169–178. doi: 10.1007/978-3-642-01793-3_18

Goertzel, B. (2007). Human-level artificial general intelligence and the possibility of a technological singularity: a reaction to ray Kurzweil's the singularity is near, and McDermott's critique of Kurzweil. *Artif. Intell.* 171, 1161–1173. doi: 10.1016/j.artint.2007.10.011

Goertzel, B., and Pennachin, C. (2007). *Artificial General Intelligence*. Vol. 2. Berlin; Heidelberg: Springer.

Google (2020). *What If Tool*. Available online at: https://pair-code.github.io/what-if-tool/

Greene, T. (2020). *2010–2019: The Rise of Deep Learning*. The Next Web. Available online at: https://thenextweb.com/artificial-intelligence/2020/01/02/2010-2019-the-rise-of-deep-learning/

Guo, H., Zheng, K., Fan, X., Yu, H., and Wang, S. (2019). "Visual attention consistency under image transforms for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 729–739. doi: 10.1109/CVPR.2019.00082

Hagendorff, T. (2019). The ethics of AI ethics-an evaluation of guidelines. *arXiv* 1903.03425.

Herschel, R., and Miori, V. M. (2017). Ethics & big data. *Technol. Soc.* 49, 31–36. doi: 10.1016/j.techsoc.2017.03.003

Hu, C., Jiang, J., and Wang, Z. (2019). Decentralized federated learning: a segmented gossip approach. *arXiv* 1908.07782.

Huszár, F. (2015). *Accuracy vs Explainability of Machine Learning Models*. inFERENCe. Available online at: http://bit.ly/2GAfW7c

Hutchinson, B., and Mitchell, M. (2019). "50 years of test (un) fairness: lessons for machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY: ACM), 49–58. doi: 10.1145/3287560.3287600

Ikuta, K., Ishii, H., and Nokata, M. (2003). Safety evaluation method of design and control for human-care robots. *Int. J. Robot. Res.* 22, 281–297. doi: 10.1177/0278364903022005001

Jan, T. G. (2020). "Clustering of tweets: a novel approach to label the unlabelled tweets," in *Proceedings of ICRIC 2019* (Jammu: Springer), 671–685. doi: 10.1007/978-3-030-29407-6_48

Jha, S., Raj, S., Fernandes, S., Jha, S. K., Jha, S., Jalaian, B., et al. (2019). "Attribution-based confidence metric for deep neural networks," in *Advances in Neural Information Processing Systems* (Vancouver), 11826–11837.

Johnson, H. (2015). *Digging Up Dark Data: What Puts IBM at the Forefront of Insight Economy*. Silicon Angle. Available online at: https://siliconangle.com/2015/10/30/ibm-is-at-the-forefront-of-insight-economy-ibminsight/

Johnston, M., Cohen, P. R., McGee, D., Oviatt, S. L., Pittman, J. A., and Smith, I. (1997). "Unification-based multimodal integration," in *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics* (Madrid: Association for Computational Linguistics), 281–288. doi: 10.3115/979617.979653

Jørgensen, S. H., Baird, A. E., Juutilainen, F. T., Pelt, M., and Højholdt, N. C. (2018). [multi'vocal]: reflections on engaging everyday people in the development of a collective non-binary synthesized voice. *ScienceOpen Res.* doi: 10.14236/ewic/EVAC18.41

Kahani, M., and Beadle, H. (1997). Decentralised approaches for network management. *ACM SIGCOMM Comput. Commun. Rev.* 27, 36–47. doi: 10.1145/263932.263940

Kendall, A., and Cipolla, R. (2016). "Modelling uncertainty in deep learning for camera relocalization," in *2016 IEEE International Conference on Robotics and Automation (ICRA)* (Stockholm: IEEE), 4762–4769. doi: 10.1109/ICRA.2016.7487679

Keren, G., Cummins, N., and Schuller, B. (2018). Calibrated prediction intervals for neural network regressors. *IEEE Access* 6, 54033–54041. doi: 10.1109/ACCESS.2018.2871713

Khan, N., Naim, A., Hussain, M. R., Naveed, Q. N., Ahmad, N., and Qamar, S. (2019). "The 51 v's of big data: survey, technologies, characteristics, opportunities, issues and challenges," in *Proceedings of the International Conference on Omni-Layer Intelligent Systems* (Crete: ACM), 19–24. doi: 10.1145/3312614.3312623

Koene, A. (2017). Algorithmic bias: addressing growing concerns [leading edge]. *IEEE Technol. Soc. Mag.* 36, 31–32. doi: 10.1109/MTS.2017.2697080

Krizhevsky, A., Nair, V., and Hinton, G. (2009). *CIFAR-10*. Toronto: Canadian Institute for Advanced Research.

LeCun, Y., and Cortes, C. (2010). *MNIST Handwritten Digit Database*.

Lee, H.-S. (2002). Optimal consensus of fuzzy opinions under group decision making environment. *Fuzzy Sets Syst.* 132, 303–315. doi: 10.1016/S0165-0114(02)00056-8

L'heureux, A., Grolinger, K., Elyamany, H. F., and Capretz, M. A. (2017). Machine learning with big data: challenges and approaches. *IEEE Access* 5, 7776–7797. doi: 10.1109/ACCESS.2017.2696365

Liu, A., Xu, N., Nie, W., Su, Y., Wong, Y., and Kankanhalli, M. S. (2017). Benchmarking a multimodal and multiview and interactive dataset for human action recognition. *IEEE Trans. Cybern.* 47, 1781–1794. doi: 10.1109/TCYB.2016.2582918

Liu, H., Feris, R. S., and Sun, M. (2011). *Benchmarking Datasets for Human Activity Recognition*. New York, NY: Springer.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., and Puri, R. (2019). "Bias mitigation post-processing for individual and group fairness," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton: IEEE), 2847–2851. doi: 10.1109/ICASSP.2019.8682620

Luo, Y., Jin, H., and Li, P. (2019). "A blockchain future for secure clinical data sharing: a position paper," in *Proceedings of the ACM International Workshop on Security in Software Defined Networks & Network Function Virtualization* (Richardson, TX: ACM), 23–27. doi: 10.1145/3309194.3309198

Marnau, N. (2019). *Comments on the "Draft Ethics Guidelines for Trustworthy AI" by the High-Level Expert Group on Artificial Intelligence*. Westminster: European Commission.

Mathews, M., Robles, D., and Bowe, B. (2017). *Bim+ Blockchain: A Solution to the Trust Problem in Collaboration*? Dublin: Dublin Institute of Technology.

Meer, P., Stewart, C. V., and Tyler, D. E. (2000). Robust computer vision: an interdisciplinary challenge. *Comput. Vision Image Understand.* 78, 1–7. doi: 10.1006/cviu.1999.0833

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

Miiller, Y. (1990). "Decentralized artificial intelligence," in *Decentralised AI* (Amsterdam), 3–13.

Mitchell, T. M. (1980). *The Need for Biases in Learning Generalizations*. New Brunswick, NJ: Department of Computer Science, Laboratory for Computer Science Research.

Mittelstadt, B. D., and Floridi, L. (2016). The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci. Eng. Ethics* 22, 303–341. doi: 10.1007/s11948-015-9652-2

Molnar, C. (2019). *Interpretable Machine Learning*. Munich: Lulu.com.

Montes, G. A., and Goertzel, B. (2019). Distributed, decentralized, and democratized artificial intelligence. *Technol. Forecast. Soc. Change* 141, 354–358. doi: 10.1016/j.techfore.2018.11.010

Morgan, S., and Waring, C. (2004). *Guidance on Testing Data Reliability*. Austin, TX. Available online at: http://bit.ly/2kjNgX4

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* 19, 98–100. doi: 10.1109/MRA.2012.2192811

Mou, X. (2019). *Artificial Intelligence: Investment Trends and Selected Industry Uses*. IFC. Available online at: https://bit.ly/3af5z6V

Naughton, J. (2019). *AI Is Making Literary Leaps–Now We Need the Rules to Catch Up*. The Guardian. Available online at: https://www.theguardian.com/commentisfree/2019/nov/02/ai-artificial-intelligence-language-openai-cpt2-release/

Nikolova, G., Kotev, V., Dantchev, D., and Kiriazov, P. (2018). "Basic inertial characteristics of human body by walking," in *Proceedings of The 15th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering and the 3rd Conference on Imaging and Visualization, CMBBE* (Lisbon), 26–29.

Osoba, O. A., and Welser, I. V. W. (2017). *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Santa Monica, CA: Rand Corporation.

Papadamou, K., Papasavva, A., Zannettou, S., Blackburn, J., Kourtellis, N., Leontiadis, I., et al. (2019). Disturbed youtube for kids: characterizing and detecting disturbing content on youtube. *arXiv* 1901.07046.

Parikh, R. B., Teeple, S., and Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. *JAMA* 322, 2377–2378. doi: 10.1001/jama.2019.18058

Price, M., and Ball, P. (2014). Big data, selection bias, and the statistical patterns of mortality in conflict. *SAIS Rev. Int. Affairs* 34, 9–20. doi: 10.1353/sais.2014.0010

Qian, K., Zhang, Z., Baird, A., and Schuller, B. (2017). Active learning for bird sounds classification. *Acta Acust. United Acust.* 103, 361–364. doi: 10.3813/AAA.919064

Radzik, L., Bennett, C., Pettigrove, G., and Sher, G. (2020). *The Ethics of Social Punishment: The Enforcement of Morality in Everyday Life*. Cambridge: Cambridge Core.

Rai, A. (2020). Explainable AI: from black box to glass box. *J. Acad. Market. Sci.* 48:137–141. doi: 10.1007/s11747-019-00710-5

Regan, P. M., and Jesse, J. (2019). Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking. *Ethics Inform. Technol.* 21, 167–179. doi: 10.1007/s10676-018-9492-2

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 1135–1144. doi: 10.1145/2939672.2939778

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., et al. (2019). "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (Nice), 3–12. doi: 10.1145/3347320.3357688

Rothwell, S., Elshenawy, A., Carter, S., Braga, D., Romani, F., Kennewick, M., et al. (2015). "Controlling quality and handling fraud in large scale crowdsourcing speech data collections," in *Proceedings of INTERSPEECH* (Dresden), 2784–2788.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., et al. (2018). Aequitas: a bias and fairness audit toolkit. *arXiv* 1811.05577.

Samek, W., Wiegand, T., and Müller, K. (2017). Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *arXiv* abs/1708.08296.

Schembera, B., and Durán, J. M. (2020). Dark data as the new challenge for big data science and the introduction of the scientific data officer. *Philos. Technol.* 33, 93–115. doi: 10.1007/s13347-019-00346-x

Schlagwein, D., Cecez-Kecmanovic, D., and Hanckel, B. (2019). Ethical norms and issues in crowdsourcing practices: a Habermasian analysis. *Inform. Syst. J.* 29, 811–837. doi: 10.1111/isj.12227

Schneider, D. F. (2020). "Machine learning and artificial intelligence," in *Health Services Research* eds J. Dimick and C. Lubitz (New York, NY: Springer), 155–168. doi: 10.1007/978-3-030-28357-5_14

Schuller, B. W., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K. R., Ringeval, F., et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH* (Lyon), 148–152.

Schütt, K. T., Gastegger, M., Tkatchenko, A., and Müller, K.-R. (2019). "Quantum-chemical insights from interpretable atomistic neural networks," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (New York, NY: Springer), 311–330. doi: 10.1007/978-3-030-28954-6_17

Setia, P. K., Tillem, G., and Erkin, Z. (2019). "Private data aggregation in decentralized networks," in *2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG)* (Istanbul: IEEE), 76–80. doi: 10.1109/SGCF.2019.8782377

Settles, B., Craven, M., and Friedland, L. (2008). "Active learning with real annotation costs," in *Proceedings of the NIPS Workshop on Cost-Sensitive Learning* (Vancouver, CA), 1–10.

Simon, M., Rodner, E., and Denzler, J. (2016). Imagenet pre-trained models with batch normalization. *arXiv* 1612.01452.

Singh, T. (2018). *Why Enterprises Need to Focus on Decentralized AI*.

Smith, C. J. (2019). "Transhumanism and distributed ledger technologies," in *The Transhumanism Handbook* ed N. Lee. (New York, NY: Springer), 529–531. doi: 10.1007/978-3-030-16920-6_34

Stahl, B., and Wright, D. (2018). Ethics and privacy in ai and big data: Implementing responsible research and innovation. *IEEE Security Privacy* 16, 26–33. doi: 10.1109/MSP.2018.2701164

Stappen, L., Baird, A., Rizos, G., Tzirakis, P., Du, X., Hafner, F., et al. (2020). *Muse 2020-The First International Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*.

Stark, T. H. (2015). Understanding the selection bias: social network processes and the effect of prejudice on the avoidance of outgroup friends. *Soc. Psychol. Q.* 78, 127–150. doi: 10.1177/0190272514565252

Sueur, C., Deneubourg, J.-L., and Petit, O. (2012). From social network (centralized vs. decentralized) to collective decision-making (unshared vs. shared consensus). *PLoS ONE* 7:e0032566. doi: 10.1371/journal.pone.0032566

Swan, M. (2015). Blockchain thinking: the brain as a decentralized autonomous corporation. *IEEE Technol. Soc. Mag.* 34, 41–52. doi: 10.1109/MTS.2015.2494358

The-European-Commission (2019). *The 2018 Reform of EU Data Protection Rules*. London: EU.

Tjoa, E., and Guan, C. (2019). A survey on explainable artificial intelligence (XAI): towards medical XAI. *arXiv* 1907.07374.

Trajanov, D., Zdraveski, V., Stojanov, R., and Kocarev, L. (2018). "Dark data in internet of things (IOT): challenges and opportunities," in *7th Small Systems Simulation Symposium* (Niš), 1–8.

Trindel, K., Polli, F., and Glazebrook, K. (2019). "Using technology to increase fairness in hiring," in *What Works*? (Amherst, MA), 30.

van Otterlo, M. (2018). Gatekeeping algorithms with human ethical bias: the ethics of algorithms in archives, libraries and society. *arXiv* abs/1801.01705.

Vellido, A., Martín-Guerrero, J. D., and Lisboa, P. J. G. (2012). "Making machine learning models interpretable," in *Proceedings of European Symposium on Artificial Neural Networks* (Bruges), 163–172.

Vidgen, B., and Yasseri, T. (2016). *P*-values: misunderstood and misused. *arXiv* abs/1601.06805. doi: 10.3389/fphy.2016.00006

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., and Ordonez, V. (2019a). "Balanced datasets are not enough: estimating and mitigating gender bias in deep image representations," in *Proceedings of the IEEE International Conference on Computer Vision* (Brighton), 5310–5319. doi: 10.1109/ICCV.2019.00541

Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S. C., et al. (2019b). "Learning unsupervised video object segmentation through visual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Ithaca, NY), 3064–3074. doi: 10.1109/CVPR.2019.00318

Wang, Y., Mendez, A. E. M., Cartwright, M., and Bello, J. P. (2019c). "Active learning for efficient audio annotation and classification with a large amount of unlabeled data," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New York, NY: IEEE), 880–884. doi: 10.1109/ICASSP.2019.8683063

Wang, Y., and Yao, Q. (2019). Few-shot learning: a survey. *arXiv* 1904.05046.

Wang, Z., Liu, K., Li, J., Zhu, Y., and Zhang, Y. (2019d). Various frameworks and libraries of machine learning and deep learning: a survey. *Archiv. Comput. Methods Eng.* 1–24. doi: 10.1007/s11831-018-09312-w

Waterhouse Cooper, P. (2019). *Responsible AI Framework*. PwC. Available online at: https://www.pwc.co.uk/services/risk-assurance/insights/accelerating-innovation-through-responsible-ai/responsible-ai-framework.html

Westphal, P., Bühmann, L., Bin, S., Jabeen, H., and Lehmann, J. (2019). *SML-Bench-A Benchmarking Framework for Structured Machine Learning*. Amsterdam: Semantic Web.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., and Yu, H. (2019). Federated learning. *Synth. Lect. Artif. Intell. Mach. Learn.* 13, 1–207. doi: 10.2200/S00960ED2V01Y201910AIM043

Yavuz, C. (2019). *Machine Bias: Artificial Intelligence and Discrimination*. (SSRN).

Zafar, S., Irum, N., Arshad, S., and Nawaz, T. (2019). "Spam user detection through deceptive images in big data," in *Recent Trends and Advances in Wireless and IoT-Enabled Networks* eds M. Jan, F. Khan, and M, Alam (New York, NY: Springer), 311–327. doi: 10.1007/978-3-319-99966-1_28

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision* (Zurich: Springer), 818–833. doi: 10.1007/978-3-319-10590-1_53

Zhang, Q., and Hua, G. (2015). "Multi-view visual recognition of imperfect testing data," in *Proceedings of the 23rd ACM International Conference on Multimedia* (Brisbane), 561–570. doi: 10.1145/2733373.2806224

Zhang, Y., Lee, R., and Madievski, A. (2001). "Confidence measure (CM) estimation for large vocabulary speaker-independent continuous speech recognition system," in *Seventh European Conference on Speech Communication and Technology* (Aalborg).

Zheng, Z., Xie, S., Dai, H.-N., Chen, X., and Wang, H. (2018). Blockchain challenges and opportunities: a survey. *Int. J. Web Grid Serv.* 14, 352–375. doi: 10.1504/IJWGS.2018.095647

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv* abs/1511.00148.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.